From Division of Neurogeriatrics
Department of Neubiology, Care Sciences and Society
Karolinska Institutet, Stockholm, Sweden

A STAR IS BORN: DEVELOPMENT OF NT* AS A NEW BIOTECHNOLOGICAL TOOL BASED ON THE MECHANISM OF PH DEPENDENT DIMERIZATION OF THE SPIDROIN N-TERMINAL DOMAIN

Médoune Sarr

Karolinska Institutet

Stockholm 2020

Cover : picture of a *Nephila clavipes* spider drinking water (credit: Anna Rising)
All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by E-Print AB 2020

ISBN 978-91-7831-671-7

A star is born: Development of NT* as a new biotechnological tool based on the mechanism of pH dependent dimerization of the spidroin N-terminal domain

THESIS FOR DOCTORAL DEGREE (Ph.D.)

Public defence: Friday 28th of February 2020 at 9.30 am.
Hörsalen (Svartsjön), Novum, floor 4, Blickagången 6, Huddinge.

By

Médoune Sarr

Principal Supervisor:
Nina Kronqvist, Ph.D.
Karolinska Institutet
Department of Neurobiology,
Care Sciences and Society
Division of Neurogeriatrics

Co-supervisors:
Prof. Jan Johansson, Ph.D.
Karolinska Institutet
Department of Neurobiology,
Care Sciences and Society
Division of Neurogeriatrics

Prof. Anna Rising, Ph.D.
Karolinska Institutet
Department of Neurobiology,
Care Sciences and Society
Division of Neurogeriatrics

Swedish University of
Agricultural Sciences
Department of Anatomy, Physiology and
Biochemistry

Opponent:
Hannes Neuweiler, Ph.D.
University of Würzburg
Department of Biotechnology and Biophysics

Examination Board:
Prof. Mikael Oliveberg, Ph.D.
Stockholm University
Department of Biochemistry and Biophysics

Prof. Amelie Eriksson Karlström, Ph.D.
Kungliga Tekniska Högskolan
Department of Protein Science
Division of Protein Engineering

Prof. Magnus Ingelman-Sundberg, Ph.D.
Karolinska Institutet
Department of Physiology and Pharmacology
Division of Pharmacogenetics

ABSTRACT

Proteins are widely used in research and in the pharmaceutical industry but the production of recombinant protein can be tedious, costly and time-consuming due to unwanted aggregation. Strategies to circumvent aggregation, e.g. the use of solubility tags, must be evaluated experimentally and successful results are not always obtained. Therefore, there is a need to develop novel solubility tags for problematic proteins. In this thesis, we propose that spiders' own solubility tag – the N-terminal domain (NT) of the spider silk proteins (spidroins) — can be harnessed to produce very aggregation-prone proteins.

NT is a pH-sensitive relay and its conformational switch from monomer to dimer determine the assembly state of the spider silk proteins. However, the molecular mechanisms of NT dimerization are unclear. In this thesis we determined which residues regulate the dimerization of NT from major ampullate spidroin and investigated whether this mechanism is conserved between distantly related NTs. Our results showed that NT dimerization requires an initial electrostatic interaction mediated by aspartate 40 with lysine 65, and that upon lowering the pH, subsequent protonation of glutamates 79, 84 and 119 result in conformational changes and stabilization. Moreover, we provided evidence that charge attraction and multistep protonation is conserved between widely different NTs but is mediated by different sets of residues.

On the basis of these findings, we designed the soluble and thermally stable NT* by disrupting the charge attraction between aspartate 40 and lysine 65. NT* was more efficient to promote solubility of aggregation-prone proteins and peptides than commonly used solubility tags. We hypothesized further that NT* would be a suitable tag to control the solubility of amyloidogenic proteins and, thereby prevent precocious aggregation. Our results support that NT* prevents aggregation of amyloidogenic proteins, allowing e.g. structural studies and identification of new anti-amyloid strategies.

This thesis presents an entirely novel approach to produce aggregation-prone recombinant proteins using the biotechnological tool NT* which was conceived on the basis of how spiders store their spidroins at high concentration without precocious aggregation.

LIST OF SCIENTIFIC PAPERS

This thesis is based on the work contained in the following papers, which are referred to by Roman numerals in the text:

I. **Sequential pH-driven dimerization and stabilization of the N-terminal domain enables rapid spider silk formation.**
Kronqvist N., Otikovs M., Chmyrov V., Chen G., Andersson M., Nordling K., Landreh M., **Sarr M**., Jörnval H., Wennmalm S., Widengen J., Meng Q., Rising A., Otzen D., Knight S.D., Jaudzems K., Johansson J.
Nature Communications. (2014); 5:3254.

II. **Conserved properties of spider silk protein N-terminal domain in spite of extensive sequence divergence.**
Sarr M., Kitoka K., Kaldmäe M., Walsh-White K.A., Landreh M., Jaudzems K., Rising A., Johansson J., Kronqvist N.
Manuscript.

III. **Efficient protein production inspired by how spiders make silk**.
Kronqvist N., **Sarr M**., Lindqvist A., Nordling K., Otikovs M., Venturi L., Pioselli B., Purhonen P., Landreh M., Sjöberg L., Robinson C., Pelizzi N., Jörnvall H., Hebert H., Jaudzems., Curstedt T., Rising A., Johansson J.
Nature Communications. (2017); 8:15504.

IV. **A spidroin-derived solubility tag enables controlled aggregation of a designed amyloid protein.**
Sarr M., Kronqvist N., Chen G., Aleksis, R., Purhonen P., Hebert H., Jaudzems K, Rising A., Johansson J.
The FEBS Journal. (2018); 85(10):1873-1885.

V. **High-yield production of amyloid-β peptide enabled by a customized spider silk domain.**
Abelein A., Chen G., Kitoka K., Aleksis R., Oleskovs F., **Sarr M**., Landreh M., Pahnke J., Nordling K., Kronqvist N., Jaudzems K., Rising A., Johansson J., Biverstål H.
Scientific Reports. (2020); 10:235.

Papers not included in the thesis:

High intracellular stability of the spidroin N-terminal domain in spite of abundant amyloidogenic segments revealed by in-cell hydrogen/deuterium exchange mass spectrometry.
Kaldmäe M., Leppert A., Chen G., **Sarr M**., Sahin C., Nordling K., Kronqvist N., Gonzalvo-Ulla M., Fritz N., Abelein A., Laín S., Biverstål H., Jörnvall H., Lane D.P., Rising A., Johansson J., Landreh M.
The FEBS Journal. (2019) doi: 10.1111/febs.15169.

Lipids Shape the Electron Acceptor-Binding Site of the Peripheral Membrane Protein Dihydroorotate Dehydrogenase.
Costeira-Paulo J., Gault J., Popova G., Ladds M.J.G.W., van Leeuwen I.M.M., **Sarr M**., Olsson A., Lane D.P., Laín S., Marklund E.G., Landreh M.
Cell Chemical Biology. (2018);25(3):309-317.

Impact of Detergents on Membrane Protein Complex Isolation.
Lee Y.C., Bååth J.A., Bastle R.M., Bhattacharjee S., Cantoria M.J., Dornan M., Gamero-Estevez E., Ford L., Halova L., Kernan J., Kürten C., Li S., Martinez J., Sachan N., **Sarr M**., Shan X., Subramanian N., Rivera K., Pappin D., Lin S.H.
Journal of Proteome Research. (2018);17(1):348-358.

CONTENTS

LIST OF ABBREVIATIONS

Aβ	Amyloid-β
BCA	Bicinchoninic acid assay
CCK58	Cholecystokinin
CD	Circular dichroism
CT	C-terminal domain
ESI	Electron spray ionization
FlSp	Flagelliform spidroin
GST	Glutathione-S-transferase
HDX	Hydrogen deuterium exchange
HSQC	Hetero single quantum coherence
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl beta-D-1-thiogalactopyranoside
MaSp	Major ampullate spidroin
MBP	Maltose binding protein
MiSp	Minor ampullate spidroin
MS	Mass spectrometry
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NT	N-terminal domain
Nus A	N-utilization substance
PGB1	Protein G B1 domain
RMSD	Root-mean-square deviation
SEC	Size exclusion chromatography
SP	Surfactant protein
SUMO	Small ubiquitin modifier
ThT	Thioflavin T
Trx	Thioredoxin
TuSp	Tubulliform spidroin
Wt	Wild-type

1 INTRODUCTION

Biotechnological research applications which are protein-based require that proteins are produced at a low cost and at a large scale. Successful protein production necessitates to find conditions that ensure a good expression level, an easy purification process and a prolonged shelf-life of the protein. At the end of this route, the proteins must be soluble, stable and correctly folded to perform their biological activities, while failure to find the right conditions would result in aggregated and inactive proteins that may be a potential liability for health if the protein is to be used for medical applications. Due to the diversity of proteins, there is no generic methodology to meet these production requirements, and each parameter has to be tailored according to the properties of the protein of interest. A common strategy to ensure a successful production is the use of solubility-enhancing tags in fusion with the proteins of interest. Fusion tags are soluble proteins with chaperone-like activities and/or thermostable characteristics helping the proteins of interest to fold properly. The efficacy of a solubility tag depends on the nature of the protein of interest and the solubility-enhancing effect must be determined empirically. Some proteins are still not readily expressed and purified despite the now available array of fusion tags, and therefore, there is a need to develop solubility tags with improved properties.

In order to identify new tags, we turned to spiders as we were intrigued by the complex, but yet elegant machinery they utilize to control the solubility of their silk proteins during storage. Spider silk protein (spidroin) solubility is affected by pH and fiber formation is mediated by the monomer/homodimer interconversion of the spidroin N-terminal domain (NT). However, the molecular details of the mechanisms by which NT work are still elusive. In this thesis work, we aim to harness the NT function as a natural solubility tag to develop a biotechnological tool devoted to the production of biomedically relevant proteins that have proven to be difficult to produce recombinantly. We determined the molecular mechanisms involved in the dimerization of NT (Papers I and II). Based on these findings we designed the mutant domain NT* that successfully enhanced the production of aggregation-prone proteins with therapeutic potential (Paper III) and provided a tool to study amyloid proteins (Papers IV and V).

2 PROTEIN STRUCTURE AND FOLDING

2.1 INTRODUCTION TO PROTEINS

Almost all biological activities of cells, such as the transport of molecules, the catalysis of biochemical reactions or the maintenance of their structural integrity, are achieved by their proteins.

Proteins consist of polymerized amino acids that have a common structure with an amino group and a carboxyl group attached to the α-carbon. They are joined into long polypeptide chains by the formation of amide bonds (peptide bonds) between the carboxyl group of one amino acid and the amino group of another. The resulting polypeptide and its order of amino acid residues is referred to as the primary structure. The amino acids are classified according to the physicochemical properties of their side chains (Table 1). In the polypeptide, the hydrophobic amino acid residues participate in Van der Waals interactions and the hydrophilic residues are involved in the formation of hydrogen bonds. Ionizable residues can engage in charge-charge interactions or salt bridges when the electrostatic interaction is coupled with hydrogen bonding. Acidic and basic residues have modulable charges that depend on their environment as their pKa:s might be altered in a hydrophobic environment or in charge clusters (1, 2). Histidine is the most versatile residue under mammalian physiological conditions since its pKa value is close to physiological pH. Although most interactions are non-covalent, cysteines (Cys) can engage in disulfide bridges, a covalent bond formed by the linkage of thiol (-SH) groups from two Cys residues into an S-S bond and concomitant release of two hydrogens.

Hydrogen bonding between atoms of the backbone chain and preferred angles for the latter, leads to the formation of a specific structural arrangement called secondary structure. Two types of secondary structures have been described by Pauling & Corey (3). The α-helix has a cylindrical shape where the carbonyl of a given residue forms a hydrogen bond with the amide that is located 4 residues further in the polypeptide chain. As a result, α-helices are quite compact since the residues involved in this type of structure are close in space. In helices that are partly solvent exposed, the side chains of hydrophobic residues point towards the core of the protein whereas the side chains of hydrophilic residues are exposed to the solvent. The second type of secondary structure, β-sheets, are built from β-strands, which are stretches of residues with an extended backbone conformation and where the side chains of adjacent residues are pointing in opposite directions. Contiguous β-strands are stacked together by backbone hydrogen bonds and can involve residues relatively distant in the protein sequence. The β-sheets can be arranged either in an anti-parallel configuration or in a parallel manner.

In the protein, the secondary structure elements and also parts of the polypeptide chain that lack defined structure interact with each other via hydrogen bonds, disulfide bridges, electrostatic, hydrophobic and Van der Waals interactions and make the structure collapse, i.e. to adopt a more compact fold where the hydrophobic patches are buried in the core of the protein. This

spatial organization of the polypeptide chain gives rise to the native conformation of the protein, which is also called the tertiary structure (4). While many proteins can have the same secondary structural elements, they fold into different tertiary structure. The configuration of the tertiary structure creates recognition sites that enable specific interactions with other molecules and allow the proteins to exert their biological functions. Some proteins require the formation of larger complexes composed of several folded polypeptides (subunits) for their activity. The manner in which the subunits are arranged is described as the quaternary structure of the protein.

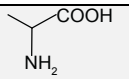
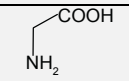
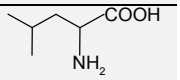
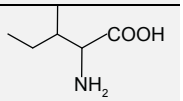
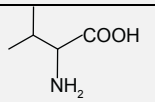
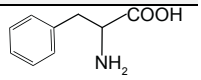
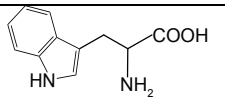
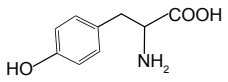
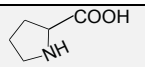
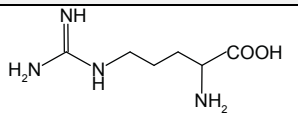
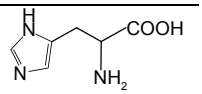
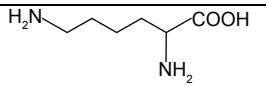
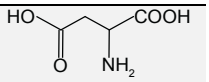
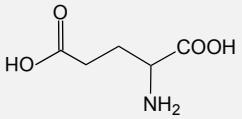
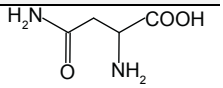
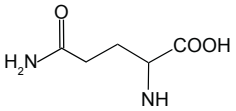
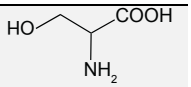
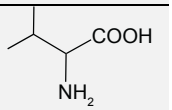
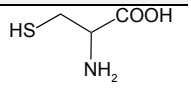
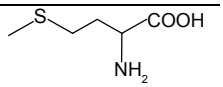
2.2 PROTEIN FOLDING

Proteins need to adopt a thermodynamically stable three-dimensional conformation, known as the native state, in order to be functional. Most proteins exist in a stable, defined, structure, but intrinsically disordered proteins lack a defined native state, and are thought to fold upon encounter of interacting molecules. The folding of proteins, *in vivo*, is co-translational (5). In addition to native contacts – i.e. interactions existing in the final three-dimensional structure of a protein – short distance and transient non-native contacts are formed during folding of a protein (6-8). These interactions make the hydrophobic parts of the polypeptide cluster in one or several partially folded intermediate structures. However, it is still unclear how the correct folding of the native state is established.

In the 1960's, Anfinsen proposed that the amino acid sequence of a protein contains all the information necessary for folding (9). This postulate was made on the basis of the spontaneous folding of ribonuclease from a denatured state and without external help. Not only the intrinsic properties of the amino acid residues composing the protein but also their distribution in the sequence is important for folding, as different pattern of polar/non-polar residues can either lead to the formation of α-helices or β-sheets (10, 11). The influence of external factors (type of solvent, pH, etc.) during assembly cannot be neglected as they can alter the structure of the final product. For example, the transmembrane and very hydrophobic lung surfactant protein (SP) C folds into an alpha helix within the lipid bilayer but forms β-sheets in aqueous solutions (12, 13).

Levinthal formulated a theory stating that proteins must take a predetermined pathway with different intermediates toward folding (14). He argued that given the high number of structures that a non-folded protein could adopt, trying out all conformations randomly would take an infinite amount of time. This theory is supported by studies of the folding process of short proteins (<100 residues) as they, indeed, fold in a fraction of a second *in vitro* (6). Moreover, the folding of small proteins is often described as a two-state transition as only one energy barrier has to be overcome (15). Transient states in a two-state folding model of small proteins are not affected by mutations which also suggest a preferential folding pathway (16).

Table 1: Summary of the 20 amino acids of the standard genetic code.

Type	Amino acid	3-letter code	1-letter code	Chemical structure
Aliphatic	Alanine	Ala	A	
	Glycine	Gly	G	
	Leucine	Leu	L	
	Isoleucine	Ile	I	
	Valine	Val	V	
Aromatic	Phenylalanine	Phe	F	
	Tryptophan	Tr	W	
	Tyrosine	Tyr	Y	
Cyclic	Proline	Pro	P	
Basic	Arginine	Arg	R	
	Histidine	His	H	
	Lysine	Lys	K	
Acidic	Aspartic acid	Asp	D	
	Glutamic acid	Glu	E	
Amine	Asparagine	Asn	N	
	Glutamine	Gln	Q	
Hydroxyl-containing	Serine	Ser	S	
	Threonine	Thr	T	
Sulfur-containing	Cysteine	Cys	C	
	Methionine	Met	M	

Another theory called the "new view" hypothesized that proteins populate various starting conformations. Therefore, proteins can take several routes and adopt diverse intermediate conformations before reaching their final native state (6, 17-20). While the proteins gain in stability as folding proceeds, the number of different low-energy intermediates to choose from becomes increasingly limited. As a result, the free-energy landscape, where the unfolded proteins navigate, is represented with a funnel shape as the energy converges to the low-energy native state (Figure 1). The walls of the hypothetical funnel are rugged because kinetic barriers have to be overcome (21, 22). Zwanzig and colleagues addressed Levinthal's paradox by speculating that energy guides the folding process (23). Even if the search for conformations is random, the proteins do not have to explore all conformations as long as only the conformations decreasing the energy are chosen down the folding routes. With this energy bias the routes to the native state take significantly less time than initially anticipated by Levinthal. Longer proteins tend to have more folding routes than short proteins and mutations can direct proteins towards preferential folding routes (16).

Despite great progress in the understanding of protein foldingit is not generally possible to predict protein structure from protein sequence only by computational methods and this remains a big challenge for protein scientists. Molecular dynamics simulations are currently limited to folding of small proteins with simple folds but they cannot accurately predict stability or thermodynamic properties of the studied proteins (6, 7).

2.3 PROTEIN AGGREGATION

The distinction between protein aggregation and protein folding can be tenuous as both pathways compete with each other and transient aggregates can be mistaken for folding intermediates (Figure 1) (24, 25). Aggregation prone sequences and/or large hydrophobic patches may be solvent exposed in folding intermediates which is energetically unfavorable. Therefore, their exposure could favor the formation of non-native intermolecular interactions leading to protein aggregation (26-29).

No general pathway for aggregations has been defined since aggregation depends on a wide range of conditions (30). The tendency to aggregate can be inherent to the protein if it has a sequence containing a high number of hydrophobic residues, forming aggregation prone regions (31). While aggregation can be driven by weak interactions only (32), proteins containing cysteines (Cys) can form covalently linked aggregates via formation of disulfide bridges (33). Furthermore, aggregation can be induced by mutations that insert hydrophobic residues within the sequence (34, 35).

Modifying the environment of the protein also affects protein folding. Molecular crowding can favor aggregation under quiescent conditions since high protein concentration increases the number of contacts between molecules (36-38). The hydrophobic residues are not the only ones playing a role in protein aggregation. Indeed, basic and acidic residues are sensitive to changes of pH or screening effects exerted by salts, and hence both pH and ionic strength can affect

aggregation (39-43). Elevated temperature can transiently result in exposure of hydrophobic patches, by promoting partial unfolding and can increase the number of collisions between molecules (44, 45). Aggregation controlled by environmental changes is generally reversible (43, 46, 47) but sometimes the process can be irreversible, especially when it causes conformational changes of the proteins (43, 48).

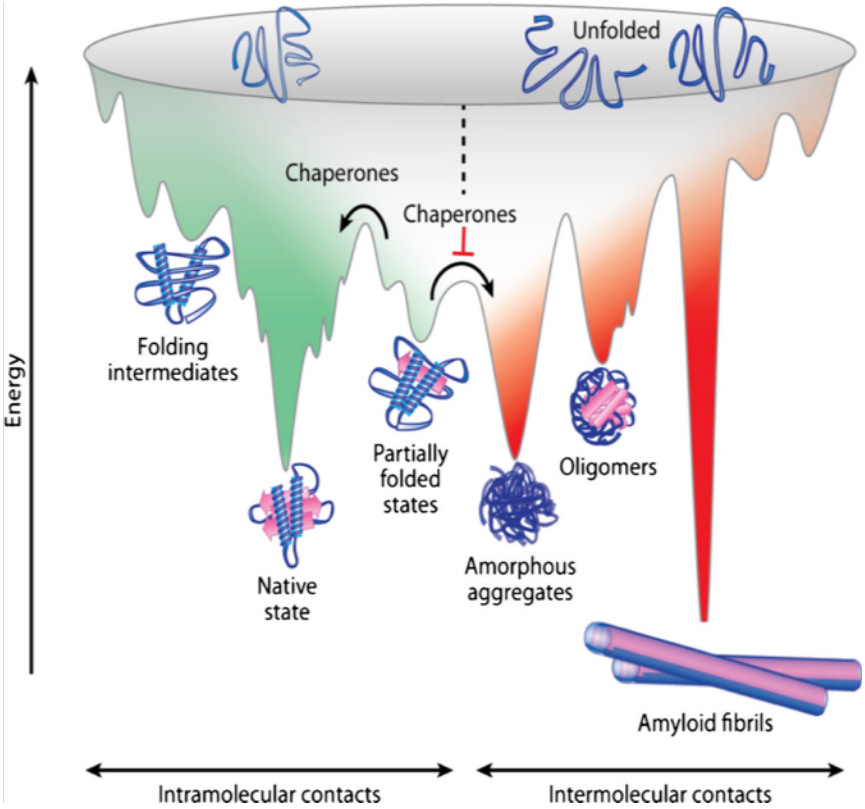


Figure 1: Schematic representation of the free energy landscape of protein folding and misfolding. Reproduced from Muntau *et al*. 2014 (49).

Irreversibly misfolded proteins represent a danger to the cells since this results in inactive proteins. They can also induce the aggregation of other proteins and mislfoded proteins as such can be cytotoxic. The cells have developed a quality control system to prevent unwanted aggregation. Protein chaperones are recruited to assist folding and protect exposed hydrophobic patches (50). When the misfolded proteins cannot be rescued, they are addressed to the ubiquitin proteasome system or the autophagy system for degradation (51, 52). However, these mechanisms of control can deteriorate and the accumulation of amorphous aggregates can provoke diseases such as cataract (53) or inclusion body myotisis (54). A specific type of particularly stable aggregates called amyloid, in which the proteins adopt a β-sheet conformation and assemble into fibrils, has been linked to several diseases (Figure 1) (55-58).

2.4 AMYLOID PROTEINS

Amyloid fibrils are proteinaceous assemblies defined by their tinctorial and ultrastructural properties and that are found in tissue deposits as well as in yeast and bacteria (59-62). Amyloid is associated with a set of diverse diseases, including transthyretin related amyloidosis, Alzheimer's disease and Parkinson disease. No curative treatments have been found for these diseases, and the current treatments can only delay the progression of the diseases (63-65). Today, 36 diseases are known to be associated with amyloid fibrils (55-58). Amyloid diseases result from protein misfolding where proteins form insoluble fibrils composed of cross-β structures (55-57, 66, 67). The lack of available treatments is probably to a large extent the result of the poor understanding of the molecular mechanisms behind the cytotoxicity observed in association to amyloid fibril formation (68, 69). It has been suggested that fragmented fibrils jeopardize the integrity of lipid membrane bilayers and can disrupt cells (70, 71). However, there is no direct correlation between the number of fibrils and the appearance of symptoms. The surface of the fibrils catalyzes secondary nucleation events that are responsible for the formation of oligomers (72, 73), which are suspected to be the main toxic species (74-81).

2.4.1 The amyloid-β peptide

Alzheimer's disease is reported to affect about 50 million people worldwide and age is the highest risk factor. Since the average population worldwide is getting older, about 130 million people are estimated to be affected by the disease in 2050 (82). The disease represents 75% of all dementia cases (83, 84). This neurodegenerative disease is characterized by the accumulation of different species of insoluble amyloid β (Aβ) proteins forming plaques in the brain (85, 86). This eventually results in memory loss and personality changes. In the quest for a better understanding of the disease on the molecular level, most attention has been focused on the 40-residue long Aβ1-40 and in particular the 42-residue long Aβ1-42, as it is the most abundant species in the plaques (87, 88), and also the most toxic Aβ species (89, 90).

Aβ1-42 has an identical sequence to Aβ1-40 except for an extra C-terminal isoleucine (Ile) and alanine (Ala) dipetide motif. They have a rather hydrophilic N-terminal part (residue 1-16) while the C-terminal part (residue 17-42) is mainly hydrophobic. Circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy revealed that monomeric Aβ1-40 and Aβ1-42 are unstructured in solution (91-94). Determination of the structures of Aβ1-40 and Aβ1-42 fibrils gave different results not only between the two petides but also between fibrils formed under different conditions and from different sources (95-99). Aβ1-40 can adopt a β-turn-β conformation with two extended anti-parallel β-strands (100). Two independent structural studies revealed that Aβ1-42 fibrils are composed of two molecules per layer and in each subfibril Aβ forms a cross-β structure shaped like a double horseshoe (101, 102). In the subfibril, the N-terminal part of Aβ is partially ordered and dynamic with a single β-strand and an unstructured part. The C-terminal part is more rigid as it contains 4 β-strands and two are interacting with each other by hydrophobic interactions. A salt bridge between the side chain

of lysine (Lys) 28 and the carboxylate of Ala 42 was also identified (98, 101, 102). This salt bridge is absent in the shorter Aβ1-40 where Lys 28 interacts with aspartate (Asp) 23 instead (96, 100). Interestingly, two studies identified a highly similar structure despite the differences in the production of the peptide (Wälti et al. used recombinant expression (102) while Colvin et al. produced the peptide by organic synthesis (101)) and in the conditions used to obtain the fibrils. Colvin and colleagues proposed that this indicates that they found the most thermodynamically stable structure of Aβ1-42 fibrils (101).

Studies revealed that Aβ1-42 has a higher rate of fibrillization than Aβ1-40 (103, 104) and no hetero fibrils were observed when both variants were co-incubated (105). The extra hydrophobic residues in Aβ1-42 compared to Aβ1-40 create a second exposed hydrophobic patch that may favor secondary nucleation and explain the higher rate of fibrillization of Aβ1-42 compared to Aβ1-40. The incompatibility of the peptides in the fibrils is supported by the structural differences between Aβ1-40 and Aβ1-42.

2.4.2 The amyloid-forming designed polypeptide β17

Despite sequence and structural differences, all proteins involved in the diverse amyloid pathologies form fibrils (55). To better understand the molecular mechanisms behind amyloidogenesis, the amyloid-like polypeptide β17 was designed (11). This *de novo* protein alternates 6 β-strands, in which the residues are organized in a polar/non-polar pattern, with 5 turns. The protein enables studies of amyloid formation as such, without interference from lost biological functions. *In vitro* studies showed that β17 is able to form amyloid fibrils (11, 106). When produced in the cytosol in eukaryotic cells, β17 fibrils promote the aggregation of mainly proteins with disordered regions, many of which are important for protein-protein interactions in signaling pathways, and newly synthesized proteins that still have not adopted their native states (106). The fibrils are then supposed to be toxic for the cells by sequestering important proteins. However, in the nucleus the toxicity is reduced and only cytosolic aggregates are shown to interfere with the nucleo-cytoplasmic transport (107). In the endoplasmic reticulum, largely unknown mechanisms, which might involve ER specific molecular chaperones, prevent β17 from fibrillating (108, 109). β17 is difficult to produce and study because of its tendency to aggregate. In *E. coli*, β17 aggregates into inclusion bodies and its purification requires the use of denaturing agents (11, 106). The instability of the purified protein, which aggregates rapidly, limits experimental studies of amyloid formation *in vitro* (108).

2.5 SPIDER SILK PROTEINS

Spider silk is one of the most remarkable polymers in nature with mechanical properties surpassing those of Kevlar and steel as this polymer is both extendable and strong (110, 111). Spider silk is also biodegradable and well tolerated when implanted *in vivo* (112, 113). These properties make spider silk a suitable candidate for various biomedical applications such as cell

culture matrices for stem cells (114, 115). The seven silk types that orb-weaving spiders use to satisfy diverse needs are described in Figure 2. The interest in producing artificial spider silk fibers is growing but the production is limited by the poor understanding of the molecular mechanisms driving fiber assembly and the limited information regarding the composition of the gland where spider silk is produced and stored.

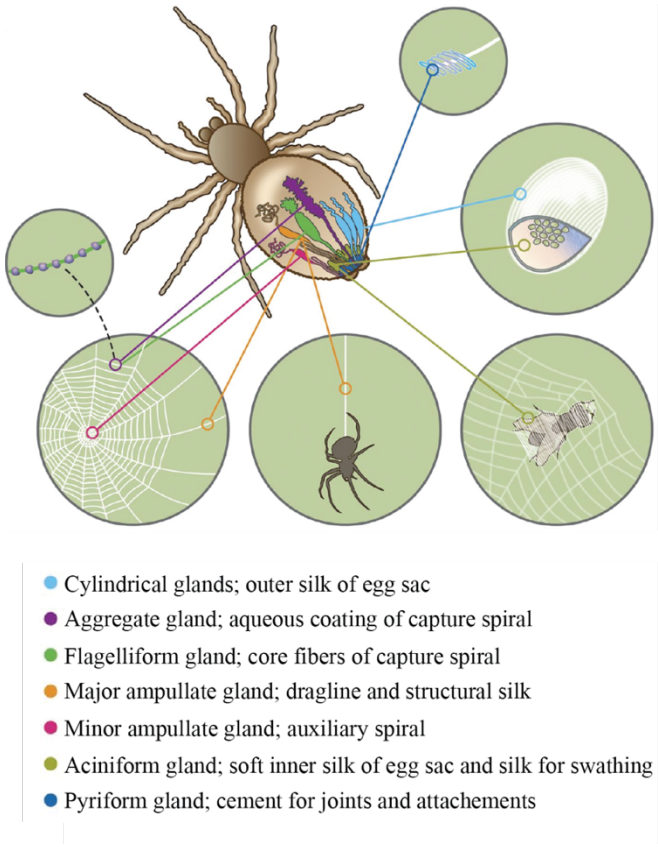


Figure 2: Illustration of female orb-weaving spider's different silk glands and the function of their respective silk fibers. Adapted with permission from Rising & Johansson: Nature Springer, Nature Chemical Biology, (116), © 2015.

Spider silk is the result of the polymerization of large proteins with variable length called spidroins. In order to respond quickly to the need of the spider, spidroins are synthesized beforehand and stored in abdominal glands which are silk specific (Figure 2). For the major ampullate spidroin (MaSp), a predicted signal peptide directs the protein to the secretory pathway in the tail of the gland (Figure 3) (117, 118). The spidroins are then stored as a liquid dope in the sac of the gland at neutral pH (119, 120), and at concentrations ranging from 30 to 50 % w/v (121, 122). However, it is still unclear how the proteins are arranged in the dope to prevent precocious aggregation. It has been suggested that spidroins adopt a micellar structure in the dope (123-125) and/or are arranged in liquid crystalline networks (126).

The silk fiber assembly occurs in the spinning duct where several factors such as pH, salt concentration and sheer forces are suspected to be fundamental in promoting assembly into

fibers (120, 127, 128). The pH is decreasing gradually along the duct, going from 7.2 at the proximal part to 5.7 midway, and reaching an unknown value, probably around pH 5, at the distal part of the duct (128-130). During the transit of spidroins in the duct, water is removed and exchange of ions occurs. All along the duct length, sodium concentration decreases whereas potassium and phosphate concentrations increase when compared to their concentrations in the proximal part of the duct (126).

2.5.1 Spidroin structure

All spidroins share a common architecture with a large repetitive central region flanked by short terminal domains (Figure 3) (131, 132). The central region of the spidroin is responsible for the mechanical properties of the silk fiber. The amino acid sequence of the central region is silk specific and gives each silk its unique mechanical properties (133). The C-terminal domain (CT) of the spidroin is relatively well conserved across silk types and spider species (134, 135). CT is an about 110 residue long and α-helical domain (128, 136-138) that form a constitutive parallel homodimer maintained mainly by hydrophobic and electrostatic interactions (136, 138, 139). In vitro, upon lowering the pH, the α-helical CT unfolds and adopts a β-sheet conformation leading to the formation of amyloid-like fibrils (128, 140). Therefore, it is hypothesized that CT act as an amyloid seed nucleating the fibril formation of the spidroins and triggering the conformational change of the repetitive parts into β-sheets. Gao et al. proposed that the linker region, between the central repetitive region and CT, acts as nucleation site (136).

2.5.2 The N-terminal domain (NT)

2.5.2.1 Function

NT is the most conserved region of the spidroin (117, 118, 131, 134), and this suggests a conserved function across silk types and species. A study on a miniature spidroin established that NT controls spidroin solubility in a pH dependent manner by preventing precocious aggregation at neutral pH (as like during storage in the gland) and by speeding up the fiber assembly at lower pH (corresponding to values observed in the duct) (141).

The underlying molecular mechanisms of the NT function were early on investigated by dynamic light scattering and X-ray crystallography, and a study suggested that NT is a homodimer at pH 7, and forms large complexes at pH 6 (141). However, further experiments using electrospray ionization mass spectroscopy, size exclusion chromatography, dynamic light scattering and analytical ultracentrifugation could not confirm the formation of large complexes (142-144) and it was later shown that NT is mainly monomeric at neutral pH and forms a dimer at lower pH (143-147).

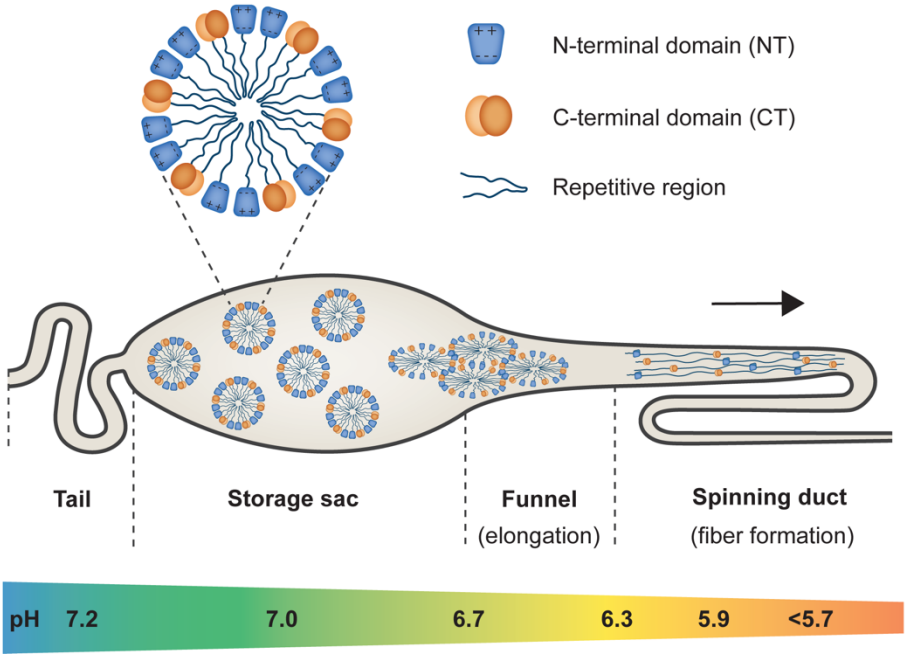


Figure 3: Illustration of the major ampullate gland describing the route of the spidroins. The spidroins are synthetized in the tail of the gland and stored in the sac conceivably in a micellar configuration at neutral pH. Assembly into fibers occurs during passage through the narrowing duct and is accompanied with a gradual decrease in pH. Reproduced from Paper III.

NT is highly soluble at high pH and can be concentrated to more than 200 mg/mL without aggregation (148) despite that two thirds of the NT sequence are predicted to have high amyloidogenic propensity (149). The association/dissociation rates of NT are independent of its concentration (145, 149). This means that in the context of the crowded dope during the storage, only ionic strength (see below) and high pH affects the solubility of NT, i.e. by keeping the domain monomeric. A decreased pH leads to dimerization and global stabilization of the dimer. The dimerization might reduce the spidroin solubility during transit in the spinning duct (142, 144, 148, 150). Interestingly, NT monomers associate at a rate close to the speed limit of diffusion and such a feature most likely enables spidroins to rapidly assemble into fibers (145). However, most experiments were carried out on isolated NT, ie without the repetitive regions, and the presence of the latter might affect the assembly process. Furthermore, experiements have usually been performed with NT concentrations that are lower than those in the gland.

2.5.2.2 Structure and molecular dynamics

Structural studies of MaSp NT show that the domain is organized as a five-helix bundle (117, 141, 143, 148, 151), and has no non-spidroin structural homologue to date (134). Structural differences are observed between the monomeric and dimeric conformation of NT. The arrangement of the helices in the monomer is dynamically flexible and incompatible with dimer formation (143, 147, 151). The homodimerization requires a rearrangement of the helices of the monomer subunits into a more compact conformation (147, 151). Moreover, the subunits in the dimer are asymmetric (151-153). The replacement of methionines (Met) in the core of

NT by leucines (Leu) results in the stabilization of monomeric NT and impairs dimerization, indicating that the plasticity of NT is driven by its methionines (147). Interestingly, a tryptophan (Trp) at position 10 is buried between helix 1 and 3 in the monomer and becomes surface exposed upon assembly (141, 142, 151). Trp 10 is a convenient probe to monitor the conformational switch of NT by fluorescence spectroscopy as the fluorescence of Trp is quenched and shifts towards a longer wavelength upon exposure to a polar solvent (141, 142, 145-147).

The pH responsiveness of NT is attributable to the dipolar distribution of its exposed acidic and basic residues. The N-terminal pole presents a cluster of positively charged amino acids while negatively charged amino acids congregate at the opposite C-terminal pole (141, 143). As a result, NT forms an antiparallel homodimer and the interactions between two subunits are mainly mediated by electrostatic interactions. The dissociation constant of the dimer is in the nanomolar range and reflects that the subunits are tightly bound in the dimer (145, 147). These interactions can be perturbed by the ionic strength since high salt concentrations (> 200 mM) delays dimerization by stabilizing monomer conformation (141-145, 147). The association rate of the monomer subunits is not affected by ionic strength but the dissociation rate increases at high salt concentration (145, 147).

NT association has been proposed to follow a selected-fit model (146). In contrast to the induced-fit model, which postulate that the binding event of a ligand to a protein induces conformational changes, the selected-fit model suggests that since proteins are dynamic, they populate a range of conformations and ligands associate with their complementary conformations (154). It cannot be excluded that both models are involved as some studies highlight a general mechanism of protein association driven by selection-fit followed by fine-tuning of the conformation after binding of the ligand (155, 156).

3 RECOMBINANT PRODUCTION AND PURIFICATION OF PROTEINS

Several organisms, both prokaryotic and eukaryotic, are today used for heterologous production of recombinant proteins and the choice of host depends on the properties of the protein and its intended use (Table 2). Eukaryotic cells, such as mammalian cells, are able to manufacture proteins with complex structures through chaperone activity and post-translational modifications but they are costly to handle and the yield of production is typically low (157). Yeast species are noteworthy alternatives due to the comparatively simple culture conditions and their ability to perform some of the post-translational modifications seen in mammalian cells, including glycosylation (158, 159). Cell-free systems do not involve living

Table 2: Overview of the expression systems commonly employed for the production of proteins (160, 161)

Expression system	Common host	Advantages	Limitations
Bacteria	*Escherichia coli*	Simple genetic manipulation Simple culture conditions High yields Cost-effective	No post-translational modifications Inclusion body formation Endotoxin contamination Unfavorable conditions for disulfide bond formation
Yeast	*Pichia pastoris* *Saccharomyces cerevisiae*	Simple genetic manipulation Post translational modifications Efficient secretion system High yields Cost-effective	Growth conditions require optimization Limited pattern of post translational modifications
Insect cells	*Spodoptera frugiperda* *Drosophila melanogaster*	Tolerate large proteins Complex post-translational modifications Medium yields	Tedious culture conditions Expensive Time consuming
Mammalian cells	Chinese hamster ovary Human embryonic kidney	Complex eukaryotic proteins Complex post-translational modifications Efficient folding machinery	Tedious culture conditions Low yields Time consuming Expensive Increased biosafety level
Cell-free	Cell lysates	Flexibility of expression conditions Incorporation of non-natural amino acids Production of toxic proteins	Background synthesis from extract use up reagents Low yields Expensive

organisms and therefore present an advantage especially for membrane proteins and toxic proteins but they still need to be improved due to the low efficiency compared to the *in vivo* systems (160, 162). *Escherichia coli* is the most commonly used prokaryotic expression system for recombinant protein production. The bacteria are easy to culture since they divide every 20 min (163), and they have a high expression yield with 1-20% of the total bacterial proteins corresponding to the recombinant proteins (164).

3.1 COMMON ISSUES WITH THE PRODUCTION OF RECOMBINANT PROTEINS

Despite the simplicity of using *E. coli* expression system, only 10 to 20% of all soluble heterologous proteins produced to date are expressed in and purified from *E. coli* (165, 166). Issues with expression and purifications of recombinant proteins are likely often unpublished, which makes it difficult to collect accurate data on this topic.

Eukaryotes and prokaryotes have different protein production machineries and heterologous production of eukaryotic proteins can be overwhelming for the *E. coli* machinery, often due to the need for chaperone activity or post-translational modifications for the proper folding of the proteins. Consequently, proteins can misfold and/or aggregate into inclusion bodies. Inclusion bodies were initially thought to contain amorphous proteins, but structural studies show that the protein structure is often dominated by a β-sheet conformation (167-169). Mitraki & King proposed that formation of inclusion bodies is the consequence of the aggregation of folding intermediates which implies that the constituent proteins might retain structural characteristics resembling the native state (170).

The disaggregation of inclusion bodies is a potential purification method since they are usually relatively pure (171). This type of aggregation is, however, not reversible by simple dilution and solubilization of inclusion bodies necessitates use of denaturing agents, such as urea or detergents (172-174). In many cases, the proteins can be refolded by lowering the concentration of the denaturing agents in a stepwise manner before or during purification. This process is labor-intensive, relatively expensive, and results in low protein yields and sometimes in inactive proteins when compared to purification under native conditions. Preventing the formation of inclusion bodies by decreasing the amount of proteins expressed is a good approach to produce soluble proteins, albeit the final yield will be considerably lower. This can be achieved by shortening the expression time, by reducing the amount of expression inducer or by lowering the temperature of expression since it slows down transcription and translation rates and gives more time for the protein to fold correctly (175-177). Another alternative to reduce protein aggregation is to co-express the target proteins with chaperones (178), or to use fusion tags (see 3.1.2).

Expression of soluble proteins is just the first step in protein production. The proteins also need to be purified and stored at conditions that prevent aggregation. Finding the adequate conditions to keep a protein soluble can be a delicate process as it is often protein dependent. Generally,

diluting the purified protein to avoid molecular crowding can keep it from aggregating (36-38). Factors like pH and ionic strength also have to be adjusted (39, 40, 42, 43).

Buffer systems provide a tight control of pH but their composition has to be chosen carefully to not provoke aggregation. For instance, phosphate buffer has been reported to promote protein aggregation (179, 180). The use of additives such as sugars and surfactants at low concentration can help to limit aggregation (181). Proteins are often stored at low temperature since high temperature promotes aggregation (44, 45) but a recent study showed that low temperature can also induce aggregation of some proteins (182). Freezing can be detrimental for protein solubility because it induces local hyper concentration of proteins and pH changes in some buffers (181, 183, 184). Moreover, repeated freeze/thaw cycles have been shown to increase aggregation (181). When the storage in solution is impractical, lyophilization can help but the process might also favor aggregation (185, 186).

3.2 SOLUBILITY TAGS

An efficient method to circumvent the formation of inclusion bodies and subsequent purification under denaturing conditions is the use of solubility tags (187-190). In this approach, the protein of interest is genetically engineered in fusion with a tag, which may enhance the expression, folding and solubility. Typically, the tag stabilizes the protein without altering its structure and function, and also facilitates downstream purification under non-denaturing conditions. Ideally, the tag should be readily removable by proteolysis and be applicable to a wide range of proteins. Various kinds of solubility tags are currently used for protein expression (Table 3). For example, the N-utilization substance A (NusA) slows down transcription by inhibiting the RNA polymerase (191). Since transcription and translation are coupled in prokaryotes, Nus A increases the solubility of its partner by slowing down the translation process, and thus, gives more time to its partner to fold properly (192). The 11.5 kDa thioredoxin (Trx) is another alternative as it might confer its thermal stability and high solubility in the *E. coli* cytoplasm to its fusion partner and also aid in the correct formation of disulfide bonds (193). Many solubility tags, Nus A and Trx included, need to be fused with small affinity tags such as a histidine tag in order to be purified. Other tags, like the maltose binding protein (MBP) or gluthathion-S-transferase (GST), possess both solubility enhancing properties and binding to specific ligands (175, 194, 195). However, GST performs poorly in *E. coli* due to its tendency to form homodimers (175). It is not clear how MBP operates but it is suspected to stabilize proteins long enough to give time for chaperones to assist folding (196). A major drawback of MBP and Nus A is that they can inflict a metabolic burden on the host expression system due to their high molecular weights.

In order to prevent interferences with the biological function or the structure determination of the protein, a protease cleavage site is engineered between the protein and the solubility tag. After successful purification of a soluble fusion protein, the tag is released proteolytically and captured by a second round of purification.

Table 3: Overview of the features of commonly used solubility tags compared to NT*. Adapted with permission from Waugh. 2005: Elsevier, Trends in Biotechnology, (188), © 2005 and Kimple et al. 2013 (190) with permission of the publisher.

Name	Origin	Organism	Size (kDa)	Advantages	Limitations	References
MaSp NT*	MaSp	*Euprosthenops australis*	13.9	Thermostable High refolding capacity Small size	Not an affinity tag	Papers III, IV
FlSp NT*	FlSp	*Nephila clavipes*	13.8	Thermostable High refolding capacity Small size Supercharged	Not an affinity tag	Papers II, V
Nus A	N-utilization substance	*Escherichia coli*	55	Efficient translation initiation Slows down translation	High metabolic burden Not an affinity tag	(192)
MBP	Maltose binding protein	*Escherichia coli*	40	Efficient translation initiation Affinity tag Mild elution conditions Cheap affinity resin May stabilize protein during folding	High metabolic burden	(197)

Name	Origin	Organism	Size (kDa)	Advantages	Limitations	References
GST	Glutathione-S-transferase	*Schistosoma japonicum*	26	Efficient translation initiation Affinity tag Mild elution conditions Cheap affinity resin	High metabolic burden Homodimer Poor solubility enhancer	(198)
Trx	Thioredoxin	*Escherichia coli*	12	Efficient translation initiation Promote disulfide bonds Thermostable Small size	Not an affinity tag	(193)
SUMO	Small ubiquitin modifier	Homo sapiens	11.5	Cleavage by SUMO protease Cleavage leave native N-terminus Small size	Endogenous SUMO proteases Not an affinity tag	(166, 199)
PGB1	Protein G B1 domain	*Steptoccocus sp*	7	Small size	Not an affinity tag	(200)

A clear advantage of the small ubiquitin modifier (SUMO), a 100 residue-long protein involved in regulation of transcription and protein transport, is that it does not require further engineering as it is efficiently recognized and cleaved by the SUMO protease (166, 201).However, the use of a SUMO tag is not applicable in eukaryotic systems as the SUMO protease is produced endogenously by these hosts. The immunoglobulin binding domain B1 from streptococcal protein G (PGB1) is advantageous compared to tags that need to be removed, as its small size (7 kDa) allows it to remain in fusion with the protein during structural studies (175).

Comparison of solubility tags are often performed on small set of proteins with a wide range of properties and it is therefore difficult to draw a general and reliable conclusion on the performance of each tag. The efficiency of a tag highly depends on the nature of the protein of interest and the solubility-enhancing effect must be determined experimentally in each case. Some proteins are still not readily expressed and purified despite the now available array of fusion tags, and therefore, there is a need to develop tags with improved properties.

4 SCOPE OF THE THESIS

The general aim of this thesis is to increase the understanding of spidroin fiber assembly and develop a biotechnological tool which addresses the struggles encountered during the production of protein-based products. The overall hypothesis is that the innovative use of the natural function of NT improves the production of aggregation-prone therapeutically relevant proteins.

More specific aims of the thesis are to:

- Increase the understanding of spidroin fiber assembly by determining the molecular mechanisms that mediate dimerization of NT (Papers I and II), and investigate if these mechanisms are conserved in distantly related NTs.

- Based on the findings in Papers I and II, harness the biological role of NT in the design of a new biotechnological tool (NT*) for expression of different aggregation-prone biological drug candidates and other biomedically relevant proteins (Papers III, IV and V).

- Investigate if NT* can be used to control the aggregation of designed and disease-relevant amyloid proteins and facilitate studies of the aggregation mechanisms (Papers IV and V).

5 METHODOLOGY

5.1 HETEROLOGOUS PRODUCTION OF PROTEINS IN *E. COLI*

5.1.1 Expression

Gräslund et al. reviewed the technical parameters to take into consideration for a successful expression and purification of proteins (202).There is a wide range of genetically engineered *E. coli* strains to choose from to produce proteins. Some strains can provide an oxidative environment inside the cells to promote the formation of disulfide bonds while others can assist the soluble expression of toxic and/or aggregation-prone proteins by co-expressing chaperone proteins or by tightly regulating expression with an inducible promoter.

Regulation of a strong promoter like the T7 promoter requires that the bacterial strains harbor a gene coding for the T7 RNA polymerase. The expression of this polymerase is under the control of the lac operon that is induced by lactose, which binds the lac repressor and detach it from the T7 promoter to make room for the T7 RNA polymerase. This induction is also accompanied by the expression of the β-galactosidase that metabolizes and depletes lactose in the expression system. The isopropyl beta-D-1-thiogalactopyranoside (IPTG) – a lactose analogue – is hence employed as an expression inducer instead. Since IPTG cannot be metabolized by the β-galactosidase, it provides a constitutive induction. Typically, protein expression is induced at a bacterial optical density of 0.5-1 at 600 nm, in order to maximize the protein yield. This range of optical density represents the mid-to-late log-phase of cell growth where the bacteria grow exponentially before entering the stationary phase. Additionally, the temperature and the duration of expression have to be optimized (see chapter 3.1.1).

5.1.2 Purification of proteins

Recombinant proteins must be isolated and purified from the endogenous proteins of *E. coli* or other expression hosts, and they must be kept in solution in order to allow subsequent experiments. In addition to adjusting the experimental conditions as described in chapter 3.1.1, experimentalists must carefully choose the methods of purification. The first step in protein purification is the extraction of proteins from the host cells. Bacteria can be lysed with chemical treatment (e.g. detergent), enzymatic treatment (e.g. lysozyme) or mechanical actions (e.g. sonication, application of high pressure) (202). The method employed to lyse the cells can affect the protein yield and purity. Moreover, the disruption of the bacterial wall releases endogenous proteases that may degrade the overexpressed proteins. The addition of protease inhibitors, during the lysis step, prevents enzymatic degradation of proteins.

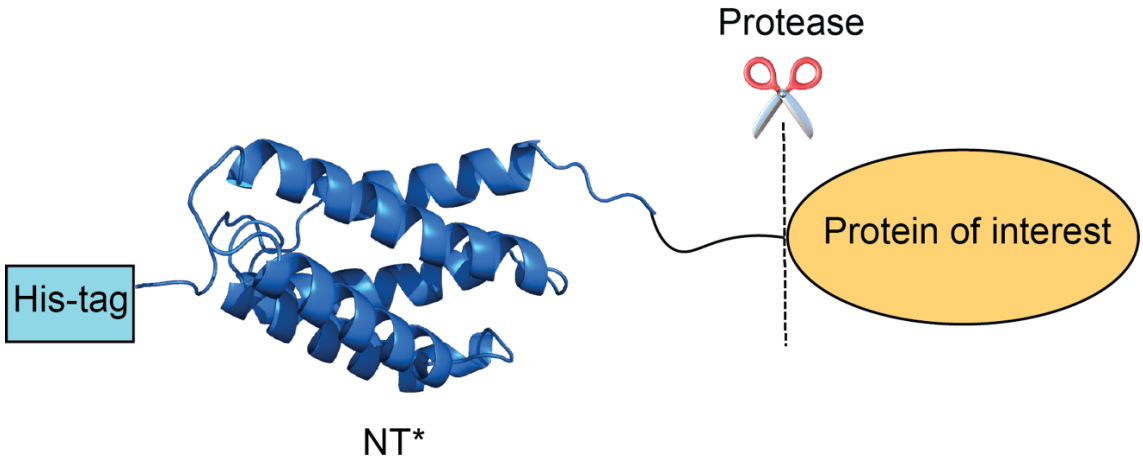


Figure 4: Schematic representation of the NT* fusion protein constructs used in Papers III, IV and V.

The second step of protein purification is to isolate the protein of interest. After lysis, the soluble proteins are separated from cell debris and insoluble proteins by centrifugation. Affinity columns like in immobilized metal affinity chromatography (IMAC) efficiently capture proteins from the centrifugation supernatant provided that they possess an affinity tag. In our experimental set-up, the constructs were designed to contain a solubility tag preceded by a His_6 affinity tag and followed by a protease recognition site before the protein of interest (Figure 4). IMAC may result in the co-purification of contaminant proteins if the level of overexpressed protein is low. This problem is addressed by submitting the sample to specific proteolysis between the solublity tag and target protein, followed by a second application of the sample to IMAC. This removes contaminants since the released target protein will pass through, while the solubility tag coupled to His_6 and the contaminants bind to the column. If the contamination persists or if the protein cannot be captured by IMAC, alternative chromatography methods can be used. These methods take advantage of the protein properties and encompass other affinity chromatography (e.g. antibodies, ligands), ion exchange (charge), size exclusion (size and shape, see chapter 5.3.1), reverse phase and hydrophobic interaction (hydropathy) chromatography. As a mean to obtain a high protein yield, each method must be fine-tuned according to the properties of the protein of interest. Janson reviewed in ref 203 the fundamentals of each chromatography methods (203).

5.2 CHARACTERIZATION OF PROTEINS BY SPECTROSCOPIC METHODS

Spectroscopic methods are used for studying the response of molecules to electromagnetic radiation stimuli. From the data generated by these methods, experimentalists can quantify proteins or infer information on the structure of a protein of interest. Spectroscopic methods were used in all the constituent papers of this thesis.

5.2.1 Absorption spectroscopy

Absorption spectroscopy, described in ref 204, achieves the quantification of protein (204). This technique exerts a direct measure of absorption based on the ability of tryptophan (Trp) and tyrosine (Tyr) to absorb the energy of a photon. The aromatic rings of their side chains act as chromophores and at 280 nm, Trp absorbs more light than Tyr. A spectrophotometer is used to measure the absorption of a sample, and consists of a light source emitting a beam that passes through a monochromator which sorts and isolates the desired wavelengths. The incident light of intensity I_0 goes through the sample and is transmitted to a detector placed in the axis of the light at an intensity $I < I_0$ as a result of the energy loss during absorption. A sample holder with neglectable absorption, such as quartz or acrylic cuvettes, must be used to ensure that the absorption observed is solely due to the sample. The absorbance A is determined by equation (i):

$$\text{(i) } A = \log_{10} \frac{I_0}{I}$$

The molar concentration C of proteins can then be calculated using the Beer-Lambert law (ii), using the path length of the sample holder (l) and the molar extinction coefficient (ε) which is specific to the molecule and depend on the type and number of aromatic residues.

$$\text{(ii) } A = \varepsilon \times l \times C$$

When Trp and Tyr are absent, the protein is "invisible" to the spectrophotometer and colorimetric methods, like the bicinchoninic acid assay (BCA), can be used to determine concentration. BCA relies on the ability of a dye (bicinchoninic acid) to detect Cu^{+} generated by the reaction of Cu^{2+} with peptide bonds. This method is highly sensitive since it can detect proteins in a range of 20-2000 μg/mL but its accuracy is limited. The assay is also relatively time-consuming as it takes at least 30 min, compared to the quasi-instantaneous results given by a spectrophotometer.

5.2.2 Fluorescence spectroscopy

Fluorescence is the emission of a photon by an excited fluorophore. As for absorbance, the fluorescent properties of proteins depend on the aromatic side chains of Trp, Tyr and to a lower extent Phe. Typically, the fluorophore in the sample is at electronic ground state (low energy level) and transitions to an excited state (high energy level) when it is hit by a photon at a wavelength of 280 nm. The excited state possesses different energy levels of vibrational states. The energy of the excited fluorophore partially dissipates by heat transfer to the solvent and the fluorophore reaches the lowest vibrational state. As the fluorophore cannot return to the ground state by losing more energy as heat, it emits a photon in the visible light spectrum, i.e. with a lower energy than the photon initially absorbed by the fluorophore. Fluorescence is measured with a spectrofluorometer, which has a similar design as a spectrophotometer except that the detector it is disposed at 90° with respect to the light source, and it includes an

additional monochromator between the sample and the detector. The principles of fluorescence spectroscopy are explained in ref 205 in greater details (205).

Fluorescence spectroscopy is particularly appropriate for proteins containing aromatic residues that conveniently relocate from a buried location into a solvent exposed one, and vice versa, upon conformational changes. The fluorescence of aromatic residues decreases and shifts to a longer wavelength (higher energy) when it transitions from a hydrophobic environment to a polar one. The high sensitivity of fluorescence to changes of polarity in the local environment of the fluorophore makes fluorescence spectroscopy a suitable tool to study protein conformational changes, stability and folding/unfolding in a dynamic system.

5.2.3 Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy, described in ref 206, mainly gives information on the overall secondary structure of proteins (206). This technique is based on the absorption of circularly polarized light by the amide group of the peptide bonds. Because proteins are chiral, the absorption of left and right circularly polarized light is unequal and the difference of polarization between incident and transmitted light is elliptic. The ellipticity spectrum in the far-UV range is characteristic of distinct secondary structures. CD is an excellent tool to monitor protein/unfolding events and conformational changes induced by the modification of the environment, and to generate kinetic and thermodynamic information. However, CD is largely a qualitative method and although the contents of secondary structure of a protein can be calculated accurately for proteins with one type of secondary structure, it is more difficult for proteins with a mixture of secondary structures.

5.2.4 Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) is together with X- ray crystallography and the recently emerging cryo-electron microscopy technique the most used method that can give structural details at the atomic level. The technique (described in ref 207) is based on the ability of atomic nuclei with a quantum spin and placed in a strong magnetic field, to respond to a weak oscillating electromagnetic field by absorbing electromagnetic energy (resonance) at different frequencies (207). The spin is an intrinsic property of nuclei with unpaired protons or neutrons and can be described by a rotation. This rotation is necessary to generate a magnetic moment as nuclei with a spin can absorb photons. The frequency of the photon is proportional to the gyromagnetic constant γ – a parameter specific to each nucleus – and the strength (B_0) of a given external magnetic field (iii).

$$\text{(iii)}\ 2\pi\nu = B_0 \cdot \gamma$$

Most NMR experiments are designed for nuclei with a spin number of ½ such as the isotopes ^{1}H, ^{13}C or ^{15}N. In the presence of an external and static magnetic field, the nuclear spin can

adopt either a ground state that is parallel to the magnetic field or a higher energy state oriented in the opposite direction of the magnetic field. At thermal equilibrium, the ground state is the most populated and in order to generate an NMR signal, the equilibrium must be disturbed. In order to do so, electromagnetic radiations with low radiofrequency, matching the energy difference between the ground state and the higher energy level, are applied to the sample for a short time. This energy transfer is called resonance, and is defined by the product of a photon frequency of oscillation (ν) with the Plank constant (h) (iv).

$$\text{(iv)}\ E = h \,.\, \nu$$

The energy difference between the two spin states is difficult to detect because it is very small, but it can be increased with a magnetic field. Linking equation (iii) and (iv) shows that the resonance of a given nucleus depends on the strength of the magnetic field (v). Consequently, the stronger the magnetic field, the bigger is the energetical transition between the two states.

$$\text{(v)}\ \Delta E = \hbar \,.\, B_0 \,.\, \gamma$$

The diversity of the electromagnetic surroundings of nuclei in a complex molecule allows to differentiate individual spins from each other. Electrons create local magnetic fields shielding the nuclei from the external magnetic field, and nuclei resonance at different frequencies in return. For instance, a proton bound to a nitrogen will have different resonance frequency than a proton bound to a carbon since the former atom is more electronegative than the latter. A difference in signal dispersion is also observed for the protons involved in different type of secondary structures. This shielding effect is called a chemical shift as the resonance frequency is shifted compared to the value for a standard molecule. The high number of nuclei in a protein renders the determination of three-dimensional structure by NMR complex due to spectral overlap in 1D-experiments. Multidimensional spectroscopy, using the NMR active nuclei of ^{1}H, ^{13}C and ^{15}N, reduces the risk of overlapping signals. For instance, the hetero single quantum coherence (HSQC) experiment is a 2D NMR spectrum, which provides a “finger print” of the protein. In an HSQC experiment, the magnetization is transferred from ^{1}H to a heteronucleus (the term refers to all atom except protons, usually ^{13}C and ^{15}N). The resulting spectrum displays one peak for each proton bound to a heteronucleus and gives indications on the fold of the protein. A folded protein has well-dispersed peaks and each residue is distinguishable but the dispersion of the spectrum becomes narrower and peaks overlap when the protein unfold.

For protein structure determination, first the spectral peaks are assigned. Then the internuclear distances are determined with experiments based on the nuclear Overhauser effect (NOE), which transfer magnetization through space. Neighboring nuclei reacting in consequence of this perturbation can, therefore be identified. Typically, nuclei within a perimeter of 5 Å exhibit a measureable NOE. The double labeling with ^{13}C and ^{15}N facilitates backbone assignments of larger proteins. From a set of spatial information, spectroscopists try to satisfy as many constraints as possible (dihedral angles, interatomic distances, etc.) during structure calculation but they can only estimate an ensemble of conformers because the distance constraints are not

precise enough. The accuracy of the structure calculation is given by the root-mean-square deviation (RMSD) of the atomic coordinates. The smaller the RMSD, the more accurate is the structure determination. The advantage of NMR spectroscopy is that NMR gives information on the dynamics of the proteins in addition to structural information.

NMR is a non-destructive and highly reproducible method provided that some conditions on the sample are met. The proteins are typically labeled during recombinant expression in minimal media supplemented with labelled sources of carbon and/or nitrogen (glucose and ammonium chloride, respectively) in order to incorporate the isotope labels. The proteins must be soluble and highly pure to prevent interferences in signal from the contaminants of the expression host. The sample is analyzed in aqueous solution, often in phosphate buffer. Alternative solutions consisting of a mixture of organic solvents, detergents or lipids, can be used for the analysis of proteins with a poor solubility in water. In order to acquire a strong signal, often protein concentrations in the millimolar range are desirable to reduce the noise-to-signal ratio. Obtaining a high concentration of aggregation-prone proteins is difficult to achieve as they will eventually fall out of solution given the long handling time necessary to record a set of spectra. For this reason, some proteins can only be analyzed in fusion with a solubility tag. Lowering protein concentration and/or temperature is also a way to overcome the aggregation issue. Although this method gives more structural information than the other spectroscopic methods described above, NMR experiments are time consuming and a set of experiment for structural determination can take up to days depending on the concentration of the sample.

5.3 EVALUATION OF PROTEIN QUATERNARY STRUCTURE AND AGGREGATION

5.3.1 Size exclusion chromatography

Size exclusion chromatography (SEC) – also called gel filtration – is both an analytical and a purification tool (203). A protein sample applied to SEC passes through a column with an inert resin composed of porous beads. The molecules in the sample penetrate the pores of different sizes and their migration is delayed to varying degrees by the passage through channels in the beads. On their pathways down the column, the retention time of the proteins is inversely proportional to their hydrodynamic radius, that is their shape and size. The shape of proteins cannot be deduced from SEC data and only an apparent molecular weight is estimated based on the calibration with a set of proteins standards. This estimation is not always accurate because the standards may have different conformations and properties compared the proteins of interest despite a similar molecular weight and, therefore, the elution profiles may differ.

Analytical SEC is employed to investigate protein size, association and aggregation. The wide range of buffers and conditions compatible with SEC allows for a good separation of monomeric proteins from larger oligomer species.

5.3.2 Mass spectrometry

Mass spectrometry is a sensitive analytical method that allows quantification and identification of molecules (208). The method can also give structural information. Mass spectrometry detects ions, separates them according to their mass-to-charge-ratio and provides their relative abundance. Molecules in solution transferred to the gas phase by electron ionization provokes extensive fragmentation of macromolecules.

Softer methods like electronspray ionization mass spectrometry (ESI-MS) were developed to overcome unsolicited fragmentation. In ESI, electrically charged droplets are generated thanks to a strong electric field that is applied on a sample in solution at the outlet of a capillary. The solvent of the droplets evaporates through their interaction with a drying gas (e.g. nitrogen) and as the concentration of charges gradually increases, the dielectric forces generated by the charge repulsion exceed the surface tension holding the droplets together. Consequently, the droplets explode and the ions are ejected into the gas phase. ESI-MS allows identification of non-covalent protein-protein interactions; however, it requires samples with high purity since contaminants can generate background noise, overlapping with the signal of the protein of interest.

Structural information can be obtained from hydrogen deuterium exchange mass spectrometry (HDX-MS) (209). This method takes advantage of the constant proton exchange between amide groups of the protein backbone and the solvent. HDX-MS detects the changes in mass ensuing the exchange of a proton from the amide backbone for a deuterium from a deuterated solvent. The rate of exchange depends on secondary structure (hydrogen bonding of the amide proton) and the accessibility of the different protein parts to solvent. For example, the residues involved in stable secondary structures or those buried in the hydrophobic core of the protein have a lower exchange rate than disordered and surface exposed residues. Thus the proton/deuterium exchange rates can give information on secondary structures, location of binding sites in protein-protein interactions, and overall protein structure and dynamics.

5.3.3 Investigating amyloid aggregation with Thioflavin T fluorescence

Thioflavin T (ThT) is a benzothiazole dye with the ability to bind amyloid aggregates and then fluoresce at 482 nm when excited at 450 nm. ThT is thought to intercalate between the side chains that are solvent-exposed and parallel to the fibril axis, and its fluorescence augments with an increased fibril mass. Due to this special feature, ThT is commonly used as a probe to monitor amyloid fibrillization kinetics *in vitro* (210).

6 RESULTS & DISCUSSION

6.1 DIMERIZATION MECHANISM OF MASP NT

The production of artificial silk-based material is appealing due its broad spectrum of potential applications. However, efforts to produce artificial silk equaling the properties of natural spider silk in a reproducible manner have so far been largely unsuccessful. This is partly due to the limited understanding of the biological events governing the assembly of spider silk proteins. For instance, the monomer/dimer interconversion of NT controls the solubility and storage of spidroins and their assembly into fibers but the detailed molecular mechanisms driving this process were largely unresolved at the start of this thesis project. In order to gain more comprehension of the underlying mechanisms of the dimerization of NT, we characterized the NT domain of MaSp from the spider *Euprosthenops australis* – herein referred to as MaSp NT. The most important findings are summarized in the next sections and a detailed account of the results are found in Paper I.

6.1.1 Structure of the dimer

Aside from the dimer structure of MaSp NT obtained by crystallography at pH 7.0 (141), the information on MaSp NT dimers were obtained with spectroscopic methods that provide details on the dynamics of the domain with respect to pH but little information on its structure. The crystal structure of the dimer is somewhat unexpected considering later reports pointing to monomers at neutral pH and dimers only at lower pH (143-147) In Paper I, structural characterization of MaSp NT by NMR spectroscopy revealed a dimer at pH 5.5 (Figure 5A) that was highly similar to the crystal dimer as reflected by the low RMSD of 1.15 Å when both structures were superimposed. On the basis of the crystal dimer structure, the Asp 40 and Glu 84 interacted in a so called handshake and it was suggested that these residues could be involved in the pH sensitive relay that controls dimerization (141) but this handshake interaction was absent from the NMR dimer structure. Possible explanations are that either the NMR data are insufficient to allow definition of the distance between the two residues, and/or the handshake interaction is transient and only detected in a conformer captured in the crystal, while it is averaged out in the NMR structure ensemble.

6.1.2 Molecular mechanism of dimerization

Identifcation of residues involved in the dimerization process of MaSp NT were based on a large set of site-directed mutations. As the transition point of dimerization was estimated at pH 6.4, a histidine (His 6) –which has a pKa around 6 – was expected to be a key residue for the process, but its deletion did not show any effect on the dimerization (141).. In a computational analysis of MaSp NT, Wallace & Shen identified Asp 39, Asp 40, Arg 60 and Lys 65 (shown

in Figure 5A) – as the residues capable of stabilizing the dimer via the formation of salt bridges (150). These conclusions were supported experimentally with mutations disrupting the salt bridge. The substitution of Arg 60 or Lys 65 for Ala destabilized the dimer (145), and the substitution of Asp 39 or Asp 40 for asparagines (Asn) prevented MaSp NT dimerization (141, 145).The presence of a salt bridge between Asp 40 and Lys 65 observed in the crystal dimer of MaSp NT was confirmed by the NMR data in Paper I. In order to better understand the role of the salt bridge, we performed site-directed mutagenesis where these residues were either substituted with neutral residues or with residues of the opposite charge. Both sets of mutations stabilized the monomer and these findings further establish the role of Asp 40 and Lys 65 in mediating the association of monomers into dimers.

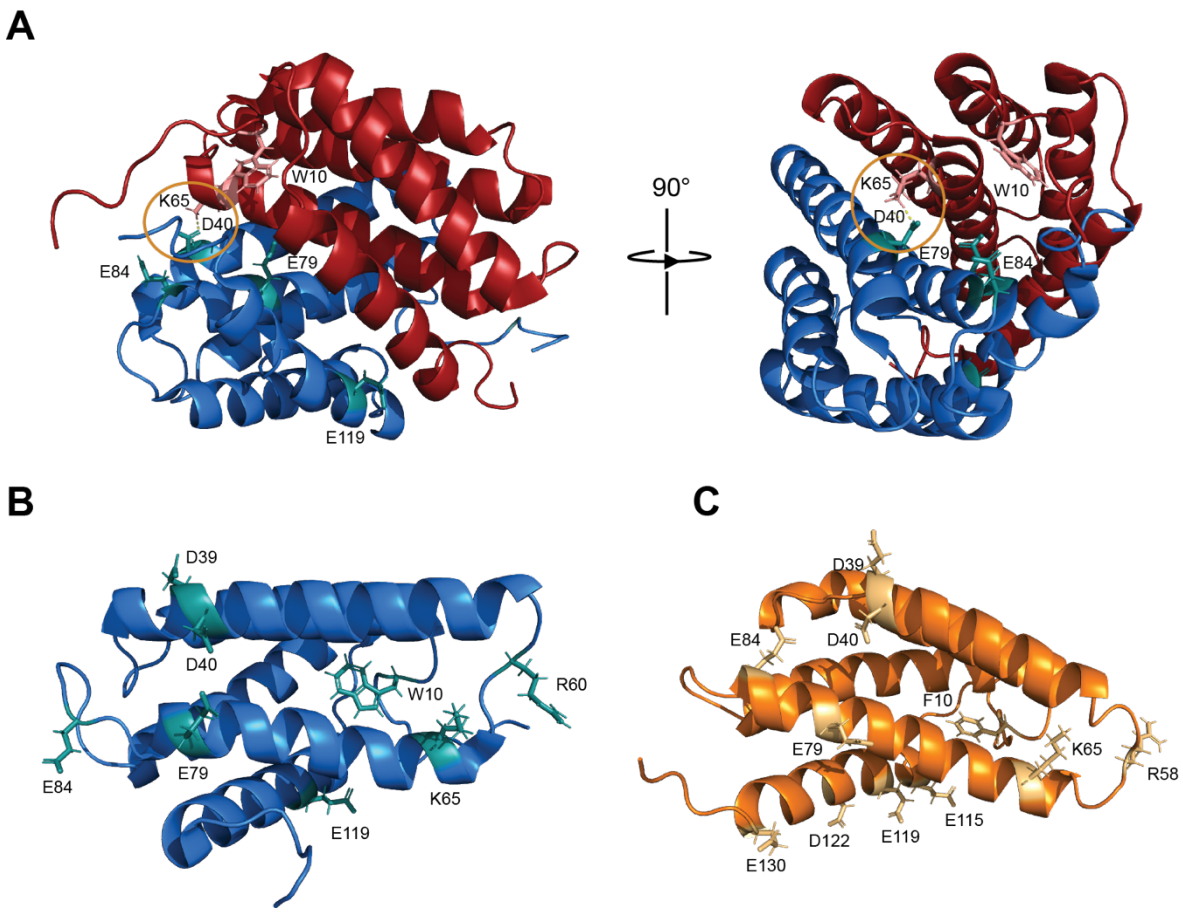


Figure 5: NMR structure of (A) MaSp NT dimer (PDB ID: 2LTH) (Paper I).The salt brigde formed by Asp 40 and Lys 65 is highlighted by an orange circle and protonated glutamates during dimerization are shown in stick representation. NMR structure of (B) MaSp NT monomer (PDB ID: 2LPJ) (151), and (C) FlSp NT monomer (accession code pending, Paper II). The charged residues and the residues involved in the dimerization process are shown in stick representation.

The cluster formation of acidic residues, which can increase the pKa values of their side chains (2), and the pH dependence of the dimerization of NT imply that titratable acidic residues are participating in the process but the experimental efforts employed so far could not identify all

the residues involved. Although they are not surface exposed, Glu 79 and Glu 119 were considered for investigations. The mutation of Glu 119 to Gln – simulating a constitutive protonated state with the amide of the side chain acting as a hydrogen donor – stabilized the dimer conformation (145), and a dimer was likewise identified by mass spectrometry at both high and low pH when the Glu 79 was replaced by Gln (144). In agreement with these results, molecular dynamic simulations singled out E79 and E119 as potential candidates for protonation (150).

In Paper I, we hypothesized that several residues are acting as a pH sensitive relay and that the effect of their protonation is cumulative. We first designed a double mutant where both Glu 79 and Glu 119 were replaced by Gln (MaSp $NT_{E79Q/E119Q}$), and we showed that the monomer/dimer equilibrium was significantly shifted towards the dimer although the dimer state was not fully reached. Therefore, another residue was thought to be involved to fully associate the monomer subunits. Glu 84 was deemed a good candidate to complete the protonation pattern because it is relatively well conserved among NTs. The dimer was further stabilized with the addition of the Glu 84 to Gln mutation (MaSp $NT_{E79Q/E84Q/E119Q}$) and similar HSQC NMR spectra were observed for the wt MaSp NT dimer and MaSp $NT_{E79Q/E84Q/E119Q}$ although the latter was analysed under conditions that favor the monomer conformation (high pH and salts). The estimation of proton uptake from the Trp fluorescence data suggested that wt MaSp NT captures 3 protons upon dimerization while MaSp $NT_{E79Q/E84Q/E119Q}$ did not pick up any proton, confirming that the dimer conformation was locked in the triple mutant. Interestingly, the protonation proceeds gradually during dimerization because the pH of Trp fluorescence interconversion is 6.5 but the last protonation occurs at pH 5.7. The shift of $NT_{E79Q/E119Q}$ towards a more dimeric population is indicative of that the titration of Glu 79 and Glu 119 is the first step of protonation required for dimerization to occur at pH 6.5. It has not been determined experimentally whether it is the protonation of Glu 79 or Glu 119 that happens first but the slightly more elevated theoretical pKa of Glu 79 (150) suggests that Glu 79 is protonated first. Their close pKa values suggest that these protonation events happen in quick succession or in parallell.

The role of Glu 84 is ambiguous because its individual mutation promoted the monomer conformation (141, 145) while the mutation had an opposite effect when it was part of MaSp $NT_{E79Q/E84Q/E119Q}$. It is possible that the charged Glu 84 at first interacts electrostatically with Arg 60 and Lys 65 to promote the interaction between two monomer subunits and the subsequent protonation of this residue in the partially stable dimer strengthens the association. An indirect effect on the dimerization also appears likely considering that the mutation of Glu 84 removes a charge in the cluster of negative charges. The charge removal might result in decreased pKa of the surrounding charged residues, lowering, *de facto*, the pH at which dimerization occurs.

In summary, we established in Paper I that the dimerization of MaSp NT is initiated by the electrostatic interaction between monomers as mediated primarily by Asp 40 and Lys 65 (Figure 6). The pKa of Glu 79 and Glu 119 becomes elevated when the two subunits are getting

closer and this leads to their protonation upon lowering the pH below 6.5. The resulting dimer is partially stable and requires the protonation of Glu 84 to be locked in a fully stable conformation. Our findings offer biologically relevant insights on how spiders tightly control the dimerization of NT to manage their silk formation. Precocious aggregation of the spidroins is avoided by keeping NT in its monomer conformation during storage and their fast assembly into fibers is promoted when NT dimerizes. *In vitro* characterizations of an NT only (ie without repetitive parts) do not necessarily reproduce its biological role in the gland where NT is connected to the large repetitive parts and present at high concentration. However, it was suggested that NT dimerization is concentration independent (145, 149), and we discovered in Paper I that the interconversion from monomer to dimer is also observed by Trp fluorescence spectroscopy on a construct where MaSp NT is in fusion with 5 repeats of the central region. Taking these results into consideration, it is tempting to speculate that NT dimerization as described here would only be marginally different in the context of the gland.

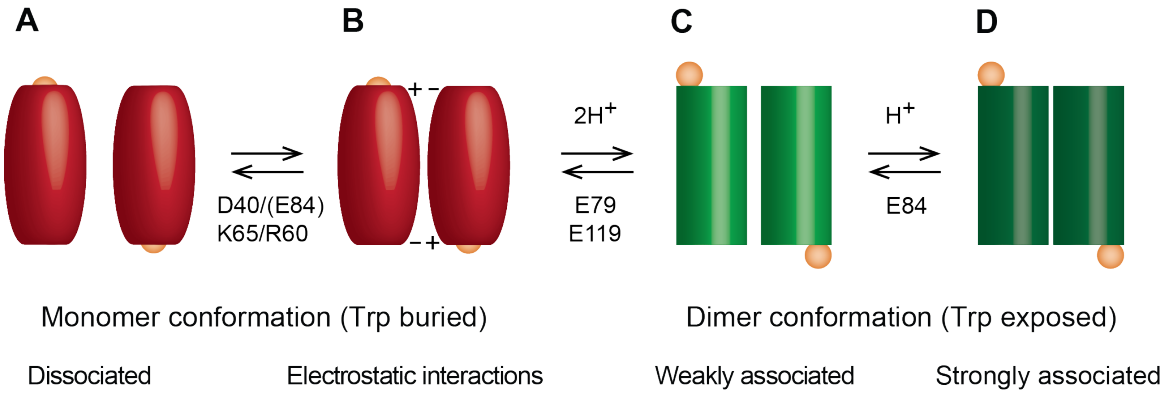


Figure 6: Proposed mechanism for MaSp NT dimerization. **(A)** MaSp NT molecules with a monomer conformation. **(B)** Electrostatic interactions mediated by D40, K65, R60 and possibly E84. **(C)** partially stable MaSp NT dimer after the protonation of E79 and E119 around pH 6.5. **(D)** Fully stable MaSp NT dimer after the protonation of E84 around pH 5.7. Reproduced from Paper I.

In the light of our results, we suggest that both micelles and a liquid crystalline state can occur during storage of spidroins in the gland (123-126). It is possible that both organizations exist at different locations of the gland. We hypothesize that the spidroins are first arranged in micelles where NT is mainly monomeric and interact with other subunits with transient electrostatic interactions. At the entrance of the duct, NT forms partially stable dimers that pre-align the spidroins, possibly in a liquid crystalline arrangement. The pre-alignment enables the rapid conversion of the spidroins into fibers when they proceed down the duct, while structural rearrangement of the spidroins still can occur, and provides a safety mechanism against precocious aggregation. As pH keeps dropping, NT dimers are further stabilized and interconnect the spidroins which allows propagation of pulling forces through the polypeptide chain and facilitates the structural conversion of the repetitive parts.

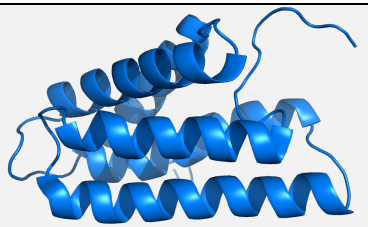
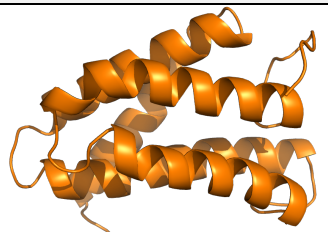
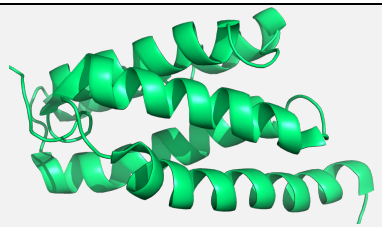
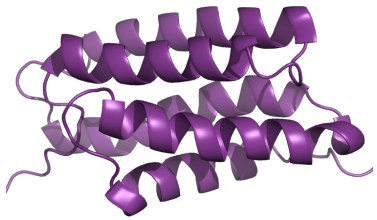
6.2 A CONSERVED MECHANISM ACROSS SPECIES AND SILK TYPES

Spiders and their silk are the product of million years of evolution (211), but yet, the NT domain of the spidroin remains highly conserved (117, 118, 131, 134). This suggests that NT has a conserved function and raises the question whether this function is mediated by the same mechanism through spider species and silk types. The NT domain of the flagelliform spidroin (FlSp) from the spider *Nephila clavipes* – herein referred to as FlSp NT – caught our attention because it shares only 35% sequence identity with the MaSp NT of *E. australis* studied in Paper I. This degree of sequence identity is so low that three-dimensional structural similarity can not be taken for granted based on sequence identity only. Moreover, FlSp NT lacks the characteristic Trp of MaSp NT at position 10, which is thought to drive the conformational switch of MaSp NT during dimerization (146). We asked ourselves what the implications of these differences are on the behavior of FlSp NT. In Paper II, we investigated the dimerization mechanism of FlSp NT to determine if it is governed by the same mechanism as MaSp NT despite their distant evolutionary relationship.

We found that wt FlSp NT form a dimer upon lowering the pH and the determined NMR structure revealed a close structural similarity to the MaSp NTs from *E. australis*, *N. clavipes* and *Latrodectus hesperus*, and MiSp NT from *Araneus ventricosus* with a three-dimensional structure organized as a dipolar 5-helix bundle (152, 212, 213). Unlike the aforementioned NTs, FlSp NT has a buried Phe at position 10 instead of a Trp. The more hydrophobic nature of Phe compared to Trp raises the question whether the helices rearrange during dimerization and expose the Phe to solvent, as for Trp in MaSp NT. In Paper II, we introduced a Trp at position 10 (NT_{Trp}) by site directed-mutagenesis. The Trp fluorescence analysis of the FlSp NT_{Trp} exhibited a sensitivity to pH demonstrated by the decrease in fluorescence amplitude at pH < 6.8. However, the response to pH was not accompanied by a major rearrangement of the helices since no shift in the fluorescence wavelength was observed. Trp fluorescence is typically quenched by neighboring protonated acidic side chains but there are no acidic groups in the vicinity of Trp 10 in FlSp NT (Figure 5C). Therefore, it is more plausible that Trp is already solvent exposed in the monomer, because it does not fit the shallow pocket normally hosting a less bulky Phe, and the dimerization provokes a structural rearrangement that moves Trp 10 close to a protonated residue. Our data indicates that having a wedged Trp swinging out from a buried position is not a prerequisite to toggle the conformation of NT during dimerization. It has been recently suggested that plasticity from the presence of Met residues drive the conformational changes of NT (147).

The absence of Trp in wt FlSp NT renders the monitoring of dimerization with Trp fluorescence spectroscopy impossible. By designing site-directed mutants previously used for the characterization of MaSp NT, we used the same methods – fluorescence spectroscopy aside – to provide evidence that the electrostatic intermolecular interaction between Asp 40 and Lys 65 that associate monomers into dimers is preserved in FlSp NT.

Table 4: Table summarizing the molecular mechanisms identified for the dimerization process of NT from different silk types and species. Asp 39 and Arg 62 in MiSp NT correspond to Asp 40 and Lys 65 of MaSp NT.

Silk type	Species	Salt bridge	Protonation	Structure	Ref
MaSp	*E. australis*	Asp 40 -Lys 65	Glu 79 Glu 84 Glu 119		Paper I, (151)
FlSp	*N. clavipes*	Asp 40-Lys 65	Not identified		Paper II
MiSp	*A. ventricosus*	Asp 39- Arg 62 Not investigated	Glu 73 Glu 79 Glu 119		(212)
MaSp	*L. hesperus*	Not investigated	Glu 79 Glu 119 Third residue was not identified		(213)

The importance of the salt bridge between these two residues has also been highlighted in the MaSp NT of *N.clavipes* (152), and Otikovs and colleagues revealed the formation of a salt bridge between Asp 39 and Arg 62 (two residues corresponding to Asp 40 and Lys 65 of MaSp NT) in the dimer of MiSp NT. Although the role of Asp 40 and Lys 65 – and corresponding residues in MiSp NT – was not explored for the other characterized NTs, the residues are most likely as important in these NTs since the charges are well conserved at these positions.

The characterization of FlSp $NT_{E79Q/E84Q/E119Q}$ in Paper II provided partly conflicting results. SEC suggested that FlSp $NT_{E79Q/E84Q/E119Q}$ undergoes a pH sensitive dimerization but ESI-MS indicated an inability to dimerize. The NMR data could rationalize these results since a mixture of dimer intermediates together with monomers were observed, which fits nicely with the selected-fit model proposed for NT association in a previous study (146). A similar intermediate conformation was observed for MaSp $NT_{E79Q/E119Q}$ in Paper I. NMR data showed that at high pH, the conformational equilibrium was shifted towards the monomer and that low pH promoted the formation of partially stable dimers but the resolution of SEC is too low to distinguish the high populated conformational state from the low populated one. ESI-MS of FlSp $NT_{E79Q/E84Q/E119Q}$ presumably dissociates loosely associated dimer intermediates.

We were not able to identify all the residues responsible for locking the FlSp NT dimer in Paper II and the elucidation of a dimer structure will eventually shed light upon the uncertainties regarding what residues are protonated. The presence of dimer intermediates in $NT_{E79Q/E84Q/E119Q}$ illustrates that at least one of the Glu considered herein, possibly Glu 79 due to its strict conservation across all NTs, is implicated in protonation events. Both Glu 84 and Glu 119 might not be involved in the dimerization of FlSp NT and the orientation of the side chains located at the dimer interface in the monomer structure points to Glu 115, Glu 130 or Asp 122 as potential candidates for titration (Figure 5C). Interestingly, these residues have no corresponding residues in other studied NTs. In MaSp NT, Glu 84 is located in a loop and has, therefore, more flexibility to perform a dual role (handshake and electrostatic attraction between subunits), while in FlSp NT it is positioned at the end of helix 3 with less freedom to move. The molecular mechanisms of dimerization of MiSp NT from *A. ventricosus* and MaSp NT from *L. hesperus* are comparable to FlSp NT and MaSp NT (*E. australis*) with the implication of Glu 79 and Glu 119 in protonation events (Table 4) (212, 213). Nevertheless, Glu 84 does not play a role in the pH-dependent interconversion of MaSp NT from *L. hesperus* and the exact residue for its replacement has still not been discovered (213), whereas Glu 84 is simply absent in MiSp and its role is assumed by Glu 73 instead (212). In the primary structures of TuSp NTs, Glu 119 is replaced by a basic residue and no acidic groups are present near this position. For this reason, TuSp NTs are intriguing because they must involve a new set of residues to protonate for dimerization, provided that it takes place at all.

Altogether, our findings validate a mechanism of dimerization of FlSp NT comparable to other studied NTs where electrostatic interactions followed by protonation events stabilize the dimer formation (Table 4), but it involves a new set of residues that are yet to be identified. Therefore, the role of NT as a pH sensitive relay and the overall mechanisms governing its dimerization

are likely to be conserved for most spiders. Babb et al reported that all silk glands produce, albeit at lower amounts, different spidroins than the ones that make up the bulk of their silk (131). The overall similar mechanisms for pH dependent dimerisation of different NTs, but based on different sets of residues, could imply that spiders have developed a mechanism that prevents hetero-assembly of co-existing different spidroins, but at the same time ensure that they homo-dimerise in the same pH range.

6.3 NT* AS A SOLUBILITY TAG

6.3.1 Development of NT*

NT plays an important role in keeping the spidroins soluble at high concentration during storage in the gland. We wondered whether it is possible to exploit the solubility enhancing competence of NT for the development of biotechnological applications since protein aggregation is a major challenge for the elaboration of protein-based therapeutic approaches. The biotechnological tools and methods developed to address protein aggregation are not universal and need to be widened. Our investigations in Paper I and II showed that NT has a high thermal stability and folding capacity. These characteristics are desirable for the design of solubility tags. Knowing that NT enhances solubility in its monomer conformation, we investigated ways to disrupt the salt bridge and charge attractions between Asp 40 and Lys 65 since it is a prerequisite for the association of NT into dimers.

In Paper III, we designed the double mutant MaSp $NT_{D40K/K65D}$, denoted MaSp NT*. We exchanged the position of D40 and K65 and characterized the dimerization ability, structure and thermodynamic stability. Trp fluorescence spectroscopy showed that NT* does not dimerize between pH 5.0 and 8.0. SEC and NMR confirmed this result by showing chromatograms and spectra of NT*, respectively, consistent with a monomer at both pH. Denaturation experiments revealed that NT* is as stable as the wild type dimer over the whole pH-range as determined by CD. The solubility of the mutant domain is two-fold greater than the solubility of wt NT although the net charge of the domain is conserved. The constitutive monomer conformation of NT* may be due to the introduction of a positively charged residue in the NT negative pole, and vice versa, which creates a new intramolecular electrostatic interaction pattern, thereby weakening the intermolecular interactions that are required for NT dimerization and also significantly lowering the pKa of acidic groups in the cluster.

6.3.2 A general tag for the production of aggregation-prone proteins of therapeutic relevance

Inspired from how spiders store their silk proteins in a highly soluble manner, we evaluated in Paper III, the capacity of NT* to enhance the solubility of various aggregation-prone proteins

and peptides. The selected proteins, including the human cationic antimicrobial protein 18, the cholecystokinin 58 (CCK58), Aβ1-40 and Aβ1-42, are known to resist soluble expression in *E. coli*. NT* was preceded by a His_6 tag to facilitate protein purification on a Ni-sepharose column and a cleavage site was engineered between the NT* and the protein of interest in order to remove the NT* solubility tag and allow isolaton of the target protein (Figure 4).

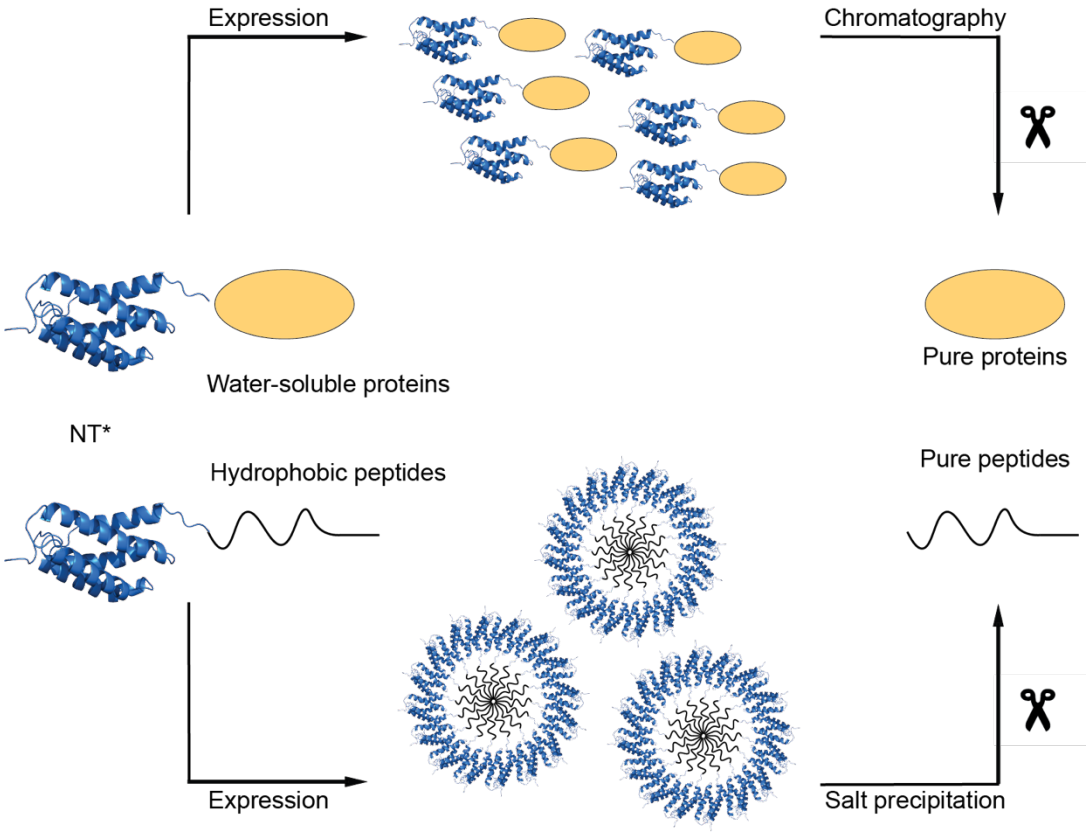


Figure 7: Scheme summarizing the purification of aggregation-prone proteins using NT* as a solubility enhancer. Following overexpression, water-soluble proteins were purified by IMAC before proteolytic release of the protein of interest. Hydrophobic peptides formed micelles in fusion with NT* and were purified by salt precipitation and pure peptides were obtained after cleavage.

Among the more thoroughly characterized proteins in Paper III was a recombinant fragment of the human pulmonary surfactant protein D (rfhSP-D). Native SP-D participates in the recognition of a wide range of pathogens and mediates immune responses by recruiting macrophages and neutrophils in the lungs (214). When previously expressed in *E. coli*, rfhSP-D was recovered from inclusion bodies with denaturing agents and at low yield (215). In fusion with NT*, rfhSP-D was expressed in a soluble manner and at higher yield than in fusion with PGB1. RfhSP-D adopted a trimeric quaternary structure characteristic of its active conformation when released from the tag.

We developed an efficient non-chromatographic method to purify transmembrane peptides and we selected the very hydrophobic rSP-C33Leu and rKL_4 as a proof of concept. The pulmonary surfactant proteins SP-B and SP-C lower the surface tension at the alveolar air-water interface, which prevents lungs from collapsing at end expiration (216-219). Native SP-C is a 4.2 kDa lipoprotein that is highly aggregation prone, and thus difficult to produce in aqueous solution, forming β-sheet aggregates and amyloid fibrils whereas it is mainly α-helical in a phospholipid environment (12, 13). Johansson and Curstedt developed SP-C33Leu – a designed analogue of native SP-C with a poly-leucine segment – which folds into a stable α-helical conformation due to the high α-helix propensity of leucine residues. In animal models of respiratory distress syndrome (RDS) – a condition engendered by prematurity or surfactant inactivation – a synthetic surfactant based on SP-C33Leu is functional (217, 220-223). A synthetic surfactant based on SP-C33Leu is currently in clinical trials for treatment of RDS in neonatal infants (224). KL_4 is an analogue of SP-B developed to treat RDS (225, 226). The peptide is 21 residues long and was designed with the motif Lys-Leu-Leu-Leu-Leu repeated four times. KL_4 is FDA approved as a component of an artificial surfactant preparation for the treatment of RDS (227, 228). Both KL_4 and SP-C proved to be water-insoluble and difficult to produce recombinantly owing to the formation of inclusion bodies.

After cell lysis and a centrifugation step to remove cell debris and other contaminants, the soluble fusion proteins were precipitated with NaCl. The subsequent pellets, pure from most *E. coli* contaminants that stay in solution, were resuspended in buffer and cleaved with cyanogen bromide to remove the solubility tag. Another salt precipitation step followed by centrifugation enabled isolation of the hydrophobic peptides which could then be resuspended in organic solvent. The procedure yielded up to 10 mg and 30 mg of pure rKL_4 and rSP-C33Leu, respectively. SEC and electron microscopy revealed that rSP-C33Leu and rKL_4 form micelle-like structures when fused with NT*. This finding is in accordance with other studies suggesting that spidroins are stored in a highly soluble manner due to their organization into micellar structures with NT shielding the hydrophobic repetitive region from the aqueous environment (123-125). The structural integrity of rSP-C33Leu was confirmed by NMR spectroscopy and the recombinant peptide was functional *in vivo* in experiments using preterm newborn rabbits as RDS models. Purified proteins from bacteria might be contaminated by lipopolysaccharides that can induce an immune response if the protein is to be injected in an animal or for pharmaceutical use (229, 230). We have not observed such an issue in our experiments with rSP-C33Leu in premature rabbit fetuses, but further purification with endotoxin columns might be required. The results are summarized in Figure 7.

The mechanisms by which NT promotes solubility remains to be established. In the same way as MBP (196), NT* might transfer its stability to the folding intermediate of the fused protein. The folding of wt NT is (60 μs) (146) and given that NT* behaves similarly it acquires its three-dimensional fold as soon as it comes out of the ribosome during translation and its high stability can protect folding intermediates of the target protein from aggregating and give them more time to fold correctly. In addition, the intermediate size of NT* (about 14 kDa) has the advantage of not being a major metabolic burden for the host.

In summary of Paper III, we effectively produced a set of highly aggregation-prone peptides and proteins, and NT* gave up to 8-fold higher amounts of functional protein than commonly used solubility tags including MBP, Trx and PGB1 (Figure 8). Moreover, we developed an elegant and efficient method for the purification of hydrophobic peptides without chromatography based on salt precipitation.

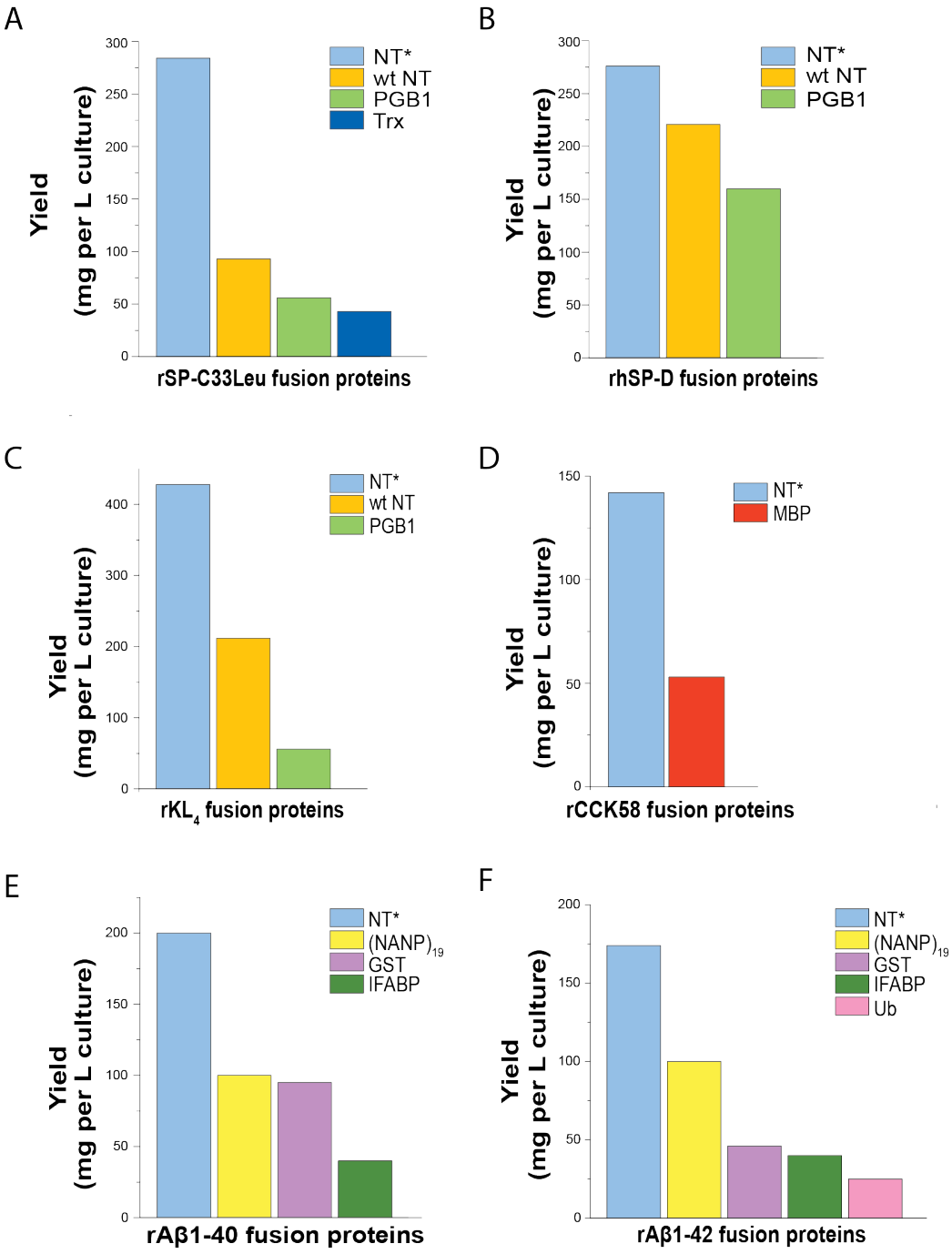


Figure 8: Comparison of protein yields obtained after IMAC purification of recombinant NT* or other solubility tags in fusion with (A) rSP-C33Leu, (B) rfhSP-D, (C) rKL_4, (D) rCCK58, (E) rAβ1-40 and (F) rAβ1-42

6.3.3 A tag to study amyloidogenic proteins

Due to their inherent pronounced tendency to aggregate, amyloid forming proteins are difficult to produce recombinantly. They are often found in inclusion bodies when overexpressed in *E. coli* and their ensuing purification is cumbersome. Furthermore, the shortage of systems exerting a tight control over their assembly jeopardizes experimental studies. In Paper IV and V, we explored the ability of NT* to control the aggregation and facilitate the study of amyloidogenic proteins.

Previous attempts to express β17 recombinantly resulted in the formation of inclusion bodies and in precocious aggregation during *in vitro* experiments (11, 106). To evaluate our system, we fused β17 to MaSp NT* in Paper IV. NT*-β17 was successfully expressed in a soluble and stable form at neutral pH but fell out of solution at pH < 6.8, limiting structural studies of the fusion protein to neutral pH or above. HSQC NMR experiments revealed that β17 was unstructured in the fusion protein and might interact with the NT* tag by elecrostatic interactions. The polypeptide was characterized by CD spectroscopy after removal of the tag by proteolysis. β17 adopted a random coil conformation mixed with β-sheet structure at pH 8.0, whereas it formed β-sheets at pH 6.0 or in presence of increasing concentrations of calcium.

Fibril formation kinetic assays using ThT as a reporter demonstrated that the fibrillation of β17 is prevented in fusion with NT*. The tag could be efficiently cleaved, which allowed the β17 protein to form ThT positive aggregates, without interference from released NT*. Nevertheless, transmission electron microscopy revealed that cleaved off β17 does not form amyloid fibrils but rather large amorphous aggregates at pH 8.0. We determined that the conversion from amorphous aggregates to fibrils only occurs at pH 6.0 or in the presence of salts. As an example, we showed that the NT* system enables interaction studies with anti-amyloid proteins, in this case Bri2 BRICHOS, to identify and evaluate potential drugs against fibril formation. The BRICHOS domain delays Aβ1-42 aggregation *in vitro* (231-234) and *in vivo* in a *Drosophila melanogaster* model (235, 236).

In Paper IV, our findings on β17 structure are in line with results from other studies on Aβ and alpha-synuclein, which suggest that monomeric amyloid proteins are disordered (91, 94, 237). The promotion of β-sheets and amyloid fibril formation by slightly acidic pH or salts is also in accordance with the literature on amyloid proteins (238-240). In this study, we developed an efficient method to produce an amyloid protein in a heterologous host which enables experimental studies without precocious aggregation.

In Paper V, we proposed a protocol to produce high amounts of Aβ1-40 and Aβ1-42 using the solubility enhancing performance of FlSp NT*. In Paper II, this mutant was developed in analogue to the MaSp NT* counterpart, from D40K/K65D mutations, as a control to study the monomer state of FlSp NT. We used it in Paper V with the rationale to investigate if this significantly more charged domain (25 charges vs 11 in MaSp NT) improves solubility enhancing properties or otherwise improves expression and purification.

The amyloid peptides were solubly expressed and could be purified with a quick and cheap purification protocol yielding higher amount of monomeric Aβ than reported in the literature (241-243). Briefly, overexpressed NT*-Aβ was purified by IMAC in the presence of 8 M urea. The urea was removed by overnight dialysis while the amyloid peptides were released from

NT* proteolytically. The mixture of tag, peptide and protease was subjected to SEC to isolate monomeric peptide. In general, NMR characterization of proteins may be hampered by the low amount of isotope-labeled proteins obtained in minimal media and the production of proteins in appropriate amounts may prove to be expensive. Furthermore the incorporation of chemically modified residues in proteins to allow for more complex NMR studies may lower the yield of production even more. Our protocol produced, in one batch, sufficient amount of isotope-labeled or fluorinated Aβ peptides to carry out NMR experiments. The strutural analysis and fibrillization kinetic assays of the recombinant Aβ1-40 and Aβ1-42 recapitulated the amyloid behavior previously described in the literature (91, 94, 232, 237, 244, 245).

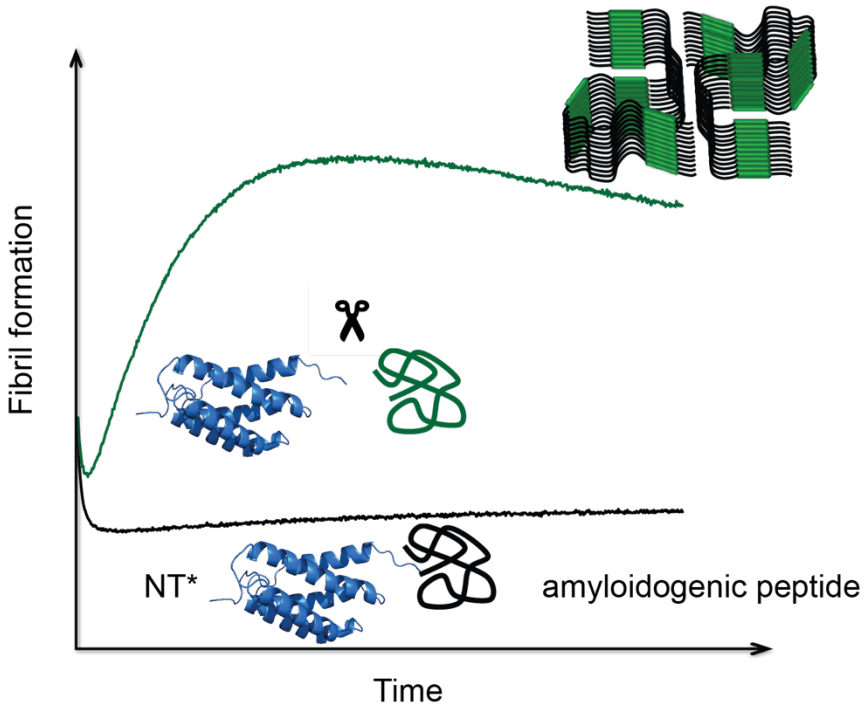


Figure 9: Representation of the tight control of NT* on the aggregation of amyloidogenic proteins. The polypeptide is largely unstructured and aggregation is inhibited in fusion with NT*. The cleavage of the fusion protein triggers the aggregation and enables controlled fibrillization studies.

The low absorbance of FlSp NT* is an advantage, as compared to the use of MaSp NT* which contains a Trp, for detection of target proteins with a low absorbance at 280 nm. We hypothesized that the greater number of charges in FlSp NT* helps it to surpass the solubility enhancing aptitude of MaSp NT*. This is supported by studies highlighting that supercharging proteins increases their solubility (246, 247). In particular, FlSp NT* might be more suitable than MaSp NT* to prevent the aggregation of amyloid proteins because FlSp is the silk type predicted to be the one structurally closest to amyloid with segments predicted to form β-hairpins (248).

The studies in Paper IV and V present a facile, fast and cheap method to recombinantly produce amyloid proteins using NT* as a solubility enhancer. Both MaSp NT* and FlSp NT* enable the production of high yields of amyloid proteins, both in rich and minimal media. The behavior of amyloidogenic proteins produced with our protocol is essentially identical to the same

proteins produced by other means. With NT*, we provide a tool which can help to gain further knowledge on the molecular mechanism underlying amyloid formation (Figure 9) NT* may also facilitate the investigation of amyloid formation of other proteins, which could aid identification of potential anti-amyloid strategies with the goal to find treatments for amyloid-related diseases.

7 CONCLUSION AND PERSPECTIVES

This thesis work sheds light on how spiders use the NT domain to manage their silk and we demonstrate how this knowledge can benefit protein production. Our studies on NT revealed that the dimerization of NT is a multistep procedure relying on electrostatic interactions followed by protonation of key residues. Using the knowledge provided by the molecular characterization of NT, we were able to design NT* and add it to the array of already available biotechnological tools which assist the production of aggregation- prone proteins.

The investigations of the molecular mechanisms behind the pH-sensitive dimerization of MaSp NT and FlSp NT (Paper I and II) shed light on the residues involved in the process. We found that for both NTs, the association is initiated by the electrostatic interaction of Asp 40 with Lys 65. In MaSp NT this interaction is stabilized by the sequential protonation of the pair Glu 79 and Glu 119 first, and the protonation of Glu 84 at lower pH. The ability of FlSp to dimerize was confirmed but the protonation events did not involve the same set of residues. Similar results has been reported for other NTs (212, 213), and taken together, the results provide evidence that the dimerization mechanism is preserved between distantly related NT. Whether the difference in titratable residues is a mechanism to prevent cross-assembly of spidroins of distinct silks remains to be determined. The feasibility of the heterodimerization can, for example, be verified by monitoring with fluorescence correlation spectroscopy the binding of NTs from two different spidroins. A possible heterodimerization would open the possibility to mix two or more recombinant silk types and make composite artificial spider silk with improved properties.

Studies of the NT domains are limited to a small amount of silk types and spider species and need to be broaden to have a complete picture of the implication of the evolution of NT on its function. For instance, several TuSp NTs are globally less charged than other NTs and they do not have some of the charged residues identified to be important for the dimerization. Future studies characterizing TuSp NT are needed to understand how this impacts NT dimerization and, *de facto*, its function.

We harnessed the dimerization mechanism of NT to design the constitutive monomer NT* and used it for biotechnological applications in Papers III, IV and V. NT* displays suitable properties for a solubility tag, namely high thermal stability and high solubility. In a comparative study in Paper III, we showed that NT* gives higher yield than commonly used tags for the production of aggregation-prone proteins and peptides that are biomedically relevant. NT* was particularly potent for the soluble expression of transmembrane peptides that were efficiently purified via a non-chromatographic method using alternate steps of salt precipitation and centrifugation. In Paper IV and V, we highlighted that NT* is an excellent tool for comprehensive studies of amyloid proteins and can help to clarify the underlying molecular mechanisms of amyloid fibril formation. For future studies, it would be interesting to include NT* in a high throughput screening for solubility tags in order to evaluate its

performance with a larger set of proteins and get more insights on its mechanism of action. In general, it would be beneficial for the scientific community to constantly report yields of purified proteins in studies involving recombinant proteins. This would help scientists to compare and evaluate the available methods for protein productions, and choose the most appropriate ones for their protein of interest.

8 ACKNOWLEDGEMENTS

I would like to sincerely thank all the persons who contributed in a way or another to the realization of this thesis.

I would like to start with the person without whom nothing would have probably happened, my main supervisor **Nina**. I cannot thank you enough for shaping the scientist that I have become. I owe you pretty much all my scientific skills. Your positive mind is contagious, so contagious that it makes things around you better. As soon as you stepped out of the building things started to go south in my experiments ☺. Despite all my catastrophies (and there were a lot of it!) you never lost hope. You always kept me grounded when I started to imagine all kind of crazy experiments. Your are really a true exemple for me and taking in consideration everything I said above, I am not sure that I am ready to let you go just yet.

I would like to express my gratitude to my co-supervisor, **Janne** for welcoming me in your wonderful group and for the guidance you provided. You are a great teacher and I learned a lot from you. In my opinion, you are simply brilliant, it seems that you never run out of ideas. You are also a great manager as you are always available to discuss the projects despite all your reponsabilities. Your dedication to research is an inspiration.

I am grateful to my co-supervisor **Anna** a.k.a spiderwoman. Your are the very first person I contacted when I wanted to join the group for my master internship. You said "yes" immediately without even checking my CV or anything else actually. Thank you for sharing your passion for spider silk, for teaching me to be rigorous in my work and for always being supportive with all my crazy ideas.

An immense thank you to **Kerstin**. When I first join the group you were introduced to me as "the most important person in the group who knows everything". I found quite soon after how true this was. You are absolutely everywhere and the amount of taks you achieve in a day is simply vertiginous. Besides your amazing professional skills, your happiness is contagious. And on top of that you love football! What more could one ask for? The transition will be difficult for the group after the Kersxit.

I would also like to say thank you:

To my great former flatmate **Axel**, our collaboration almost lasted the entire duration of my PhD. You still have to improve your Fifa skills but I must say that I admire your persistence despite your numerous defeats. To **Jessy** who joined us for the second part of my PhD and made our apartment a brighter place. And thanks to your cello for ruining my naps…

To **Oihan** my neighbor in the lab, the first witness of my catastrophies (or as you like to call them "the Médounes"-_-). It has been a pleasure to be your lab dance partner and to feature on some of your lab songs, we should definitely record an album.

To **Gefei** my other neighbor but in the office this time. Thanks for all the help you provided me. You are a mentor both in science and in football since you are the first one that I bother when I have a scientific related question or a football related question. In the past 5 years, you and me went pretty much all over Stockholm to play football with different groups.

To **Tina** for being the best student (don't start bragging!) and for all the assistance and hard work you provided in the project that must not be named again.

I would like to thank **Axel Abelein** for your kindness and patience when I had NMR related questions and **Henrik** for teaching me all the non-conventional methods to make my experiments work when the "regular" mehods did not work.

To **Juanita**, **Lorena**, **Shaffi**, **Helen** (it's good to have you back!), **Olga**, **Urmi**, **Sameer**, **Rakesh**, **Nathalie**, **Sumimala** for making the combined JJ and Rising group a place with a great atmosphere. To **Kristaps**, **Kristine**, **Margit**, **Michael**, **Pasi** for fruitful collaborations from which I learned a lot. To **Cecilia** for your availability and your kindness when I was invading your lab at SU.

To my PhD family at NVS, thanks a lot, it has been a blast being with you guys!

To **Julen** and **Daniel**, my brothers from another mother with whom I had an amazing time. Somehow and against my will, you guys managed to transform me into a fan of both Athletic and Sporting. To **Maria** for your kind heart and for being the only one who truly understands my eternal grumpiness. **Emilia** thank you for counter balancing the aforementioned grumpiness with your everlasting enthusiasm and our strange psychological discussions. To **Hazal**, my dance partners without whom it would have been impossible to get these lazy people to shake their ass when we went out. **Nuno** & **Giacomo** for showing me how to "shine like a diamond" and for being reliable food providers. To **Beeerrniiiie** for being the person I like to annoy the most. To **Ceren** for co-founding with me the club of "Fasting in Sweden is way too long, get me out of here!". To **Mona** and **Laetitia**, my French crew, for the loooong conversations although we could not communicate correctly in our mother tongue because we apparently don't master it well enough. To **Kirsten** and **Raul** for all the laughs and for teaming up with me to go to hip hop concerts. **Lea**, **Chenhong**, **Tingting**, **Luis**, **Susana**, **Gorka**, **Carmen**, **Antonio**, **Amit**, **Joana** for the nice chats and all the fun. **Maria Roos**, **Maria Ankarkrona**, **Eva** and **Helena** for your availability and for your kindness.

Mustafa, **Dina**, **Jalil** thank you for taking me under your wings and for inviting me to countless iftars. A big thank you to **Belinda** and **Joyce** for making me feel like a member of your lovely family when I just arrived in Sweden. Thank you to **Helaleh** who was always helpful and kind whenever I had something to fix in the apartment.

Merci **Magali** pour le soutien et les encouragements répétés. Merci à **Mariama**, pour les encouragements lorsqu'on était tous les deux dans la phase critique de la redaction. Thank you to **Katrine** for your enthusiasm and for preparing me to life in Sweden by teaching me the Scandinavian codes. Merci à mon amour de cousine moldavo-malienne **Fatima**, pour tout le soutien que tu m'as apporté, tu es sans aucun doute la seule raison qui ne m'a pas fait regretter d'avoir pris l'option chimie des substances naturelles en licence^^. Merci à la tout aussi adorable **Kristina** de me donner le sourire à chaque fois et pour m'emmener dans tous les coins

de Paris pour se goinfrer sans vergogne (si j'ai pris du poids maman c'est de sa faute!). Sans oublier mes deux frangins **Oussama** et **Naïm**, à qui je suis reconnaissant de leur indéfectible amitié et de leur soutien sans faille. On est ensemble!

Papa, **Maman, Mansour**, **Mame Khoudia** et tous mes **oncles**, **tantes**, **cousins** et **cousines**. Bref, **toute la famille,** je ne peux pas tous vous citer ou sinon cette partie sera plus plus épaisse que toute ma these ☺. Je vous remercie pour l'amour inconditionnel que vous m'avez porté, pour m'avoir toujours soutenu et guidé dans mes choix. Mon nom figure sur ce livre mais c'est bien à vous qu'il appartient, il est le fruit de tout votre travail <3.

9 REFERENCES

1. Flocco MM & Mowbray SL (1995) Strange bedfellows: interactions between acidic side-chains in proteins. *J. Mol. Biol.* 254(1):96-105.

2. Rostkowski M, Olsson MH, Sondergaard CR, & Jensen JH (2011) Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC Struct. Biol.* 11:6.

3. Pauling L & Corey RB (1951) The polypeptide-chain configuration in hbmoglobin and other globular proteins. *Proceedings of the National Academy of Sciences of the United States of America* 37(5):282.

4. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29(31):7133-7155.

5. Hardesty B & Kramer G (2001) Folding of a nascent peptide on the ribosome. *Prog. Nucleic Acid Res. Mol. Biol.* 66:41-66.

6. Lindorff-Larsen K, Piana S, Dror RO, & Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517-520.

7. Aumpuchin P, Hamaue S, & Kikuchi T (2019) Prediction of the initial folding sites and the entire folding processes for Ig-like beta-sandwich proteins. *Proteins: Structure, Function, and Bioinformatics*. 10.1002/prot.25862

8. Kim PS & Baldwin RL (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* 51:459-489.

9. Anfinsen CB, Haber E, Sela M, & White Jr F (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 47(9):1309.

10. Wei Y*, et al.* (2003) Stably folded de novo proteins from a designed combinatorial library. *Protein Sci.* 12(1):92-102.

11. West MW*, et al.* (1999) De novo amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci U S A* 96(20):11211-11216.

12. Szyperski T*, et al.* (1998) Pulmonary surfactant-associated polypeptide C in a mixed organic solvent transforms from a monomeric alpha-helical state into insoluble beta-sheet aggregates. *Protein Sci.* 7(12):2533-2540.

13. Gustafsson M, Thyberg J, Naslund J, Eliasson E, & Johansson J (1999) Amyloid fibril formation by pulmonary surfactant protein C. *FEBS Lett.* 464(3):138-142.

14. Levinthal C (1969) How to fold graciously. *Mossbauer spectroscopy in biological systems* 67:22-24.

15. Jackson SE & Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* 30(43):10428-10435.

16. Giri R, *et al.* (2012) Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proc Natl Acad Sci U S A* 109(44):17772-17776.

17. Bilsel O, Zitzewitz JA, Bowers KE, & Matthews CR (1999) Folding mechanism of the alpha-subunit of tryptophan synthase, an alpha/beta barrel protein: global analysis highlights the interconversion of multiple native, intermediate, and unfolded forms through parallel channels. *Biochemistry* 38(3):1018-1029.

18. Bryngelson JD, Onuchic JN, Socci ND, & Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21(3):167-195.

19. Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24(6):1501-1509.

20. Wolynes PG, Onuchic JN, & Thirumalai D (1995) Navigating the folding routes. *Science* 267(5204):1619-1620.

21. Brockwell DJ & Radford SE (2007) Intermediates: ubiquitous species on folding energy landscapes? *Curr. Opin. Struct. Biol.* 17(1):30-37.

22. Dinner AR, Sali A, Smith LJ, Dobson CM, & Karplus M (2000) Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25(7):331-339.

23. Zwanzig R, Szabo A, & Bagchi B (1992) Levinthal's paradox. *Proc Natl Acad Sci U S A* 89(1):20-22.

24. Jaenicke R (1987) Folding and association of proteins. *Progress in biophysics and molecular biology* 49(2-3):117-237.

25. Silow M & Oliveberg M (1997) Transient aggregates in protein folding are easily mistaken for folding intermediates. *Proc Natl Acad Sci U S A* 94(12):6084-6086.

26. Buchanan LE, *et al.* (2013) Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. *Proc Natl Acad Sci U S A* 110(48):19285-19290.

27. Ramakrishna D, Prasad MD, & Bhuyan AK (2012) Hydrophobic collapse overrides Coulombic repulsion in ferricytochrome c fibrillation under extremely alkaline condition. *Arch Biochem Biophys* 528(1):67-71.

28. Rashno F, *et al.* (2017) Very rapid amyloid fibril formation by a bacterial lipase in the absence of a detectable lag phase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1865(6):652-663.

29. Singh SM, Hutchings RL, & Mallela KM (2011) Mechanisms of m-cresol-induced protein aggregation studied using a model protein cytochrome c. *J. Pharm. Sci.* 100(5):1679-1689.

30. Roberts CJ (2007) Non-native protein aggregation kinetics. *Biotechnol. Bioeng.* 98(5):927-938.

31. Otzen DE, Kristensen O, & Oliveberg M (2000) Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci U S A* 97(18):9907-9912.

32. Karshikoff A (2006) *Non-covalent interactions in proteins* (World Scientific).

33. Cabra V, Vazquez-Contreras E, Moreno A, & Arreguin-Espinosa R (2008) The effect of sulfhydryl groups and disulphide linkage in the thermal aggregation of Z19 alpha-zein. *Biochim. Biophys. Acta* 1784(7-8):1028-1036.

34. Brubaker WD*, et al.* (2011) Separating instability from aggregation propensity in gammaS-crystallin variants. *Biophys. J.* 100(2):498-506.

35. Raimondi S*, et al.* (2011) Effects of the known pathogenic mutations on the aggregation pathway of the amyloidogenic peptide of apolipoprotein A-I. *J. Mol. Biol.* 407(3):465-476.

36. Alford JR, Kendrick BS, Carpenter JF, & Randolph TW (2008) High concentration formulations of recombinant human interleukin-1 receptor antagonist: II. Aggregation kinetics. *J. Pharm. Sci.* 97(8):3005-3021.

37. Schon A*, et al.* (2015) Denatured state aggregation parameters derived from concentration dependence of protein stability. *Anal. Biochem.* 488:45-50.

38. Treuheit MJ, Kosky AA, & Brems DN (2002) Inverse relationship of protein concentration and aggregation. *Pharm Res* 19(4):511-516.

39. Bhattacharya M, Jain N, & Mukhopadhyay S (2011) Insights into the mechanism of aggregation and fibril formation from bovine serum albumin. *The Journal of Physical Chemistry B* 115(14):4195-4205.

40. Bickel F*, et al.* (2016) Reversible NaCl-induced aggregation of a monoclonal antibody at low pH: Characterization of aggregates and factors affecting aggregation. *Eur J Pharm Biopharm* 107:310-320.

41. Reichert D, Groger S, & Hackel C (2017) New insights into the interaction of proteins and disaccharides-The effect of pH and concentration. *Biopolymers* 107(2):39-45.

42. Sahin E, Grillo AO, Perkins MD, & Roberts CJ (2010) Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. *J. Pharm. Sci.* 99(12):4830-4848.

43. Skamris T*, et al.* (2016) Monoclonal Antibodies Follow Distinct Aggregation Pathways During Production-Relevant Acidic Incubation and Neutralization. *Pharm Res* 33(3):716-728.

44. Kotik M & Zuber H (1992) Evidence for temperature-dependent conformational changes in the L-lactate dehydrogenase from Bacillus stearothermophilus. *Biochemistry* 31(34):7787-7795.

45. Wang W & Roberts CJ (2013) Non-Arrhenius protein aggregation. *AAPS J* 15(3):840-851.

46. Calamai M*, et al.* (2005) Reversal of protein aggregation provides evidence for multiple aggregated States. *J. Mol. Biol.* 346(2):603-616.

47. Liu J, Nguyen MD, Andya JD, & Shire SJ (2005) Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution. *J. Pharm. Sci.* 94(9):1928-1940.

48. Chi EY, *et al.* (2003) Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.* 12(5):903-913.

49. Muntau AC, Leandro J, Staudigl M, Mayer F, & Gersting SW (2014) Innovative strategies to treat protein misfolding in inborn errors of metabolism: pharmacological chaperones and proteostasis regulators. *J Inherit Metab Dis* 37(4):505-523.

50. Dahiya V & Buchner J (2019) Functional principles and regulation of molecular chaperones. *Advances in protein chemistry and structural biology* 114(1):1-60.

51. Lamark T, Kirkin V, Dikic I, & Johansen T (2009) NBR1 and p62 as cargo receptors for selective autophagy of ubiquitinated targets. *Cell Cycle* 8(13):1986-1990.

52. Wang J & Maldonado MA (2006) The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cell. Mol. Immunol* .3(4):255-261.

53. Moreau KL & King JA (2012) Protein misfolding and aggregation in cataract disease and prospects for prevention. *Trends Mol. Med.* 18(5):273-282.

54. Askanas V & Engel WK (2006) Inclusion-body myositis: a myodegenerative conformational disorder associated with Abeta, protein misfolding, and proteasome inhibition. *Neurology* 66(2 Suppl 1):S39-48.

55. Chiti F & Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* 75:333-366.

56. Benson MD, *et al.* (2018) Amyloid nomenclature 2018: recommendations by the International Society of Amyloidosis (ISA) nomenclature committee. *Amyloid* 25(4):215-219.

57. Knowles TP, Vendruscolo M, & Dobson CM (2014) The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* 15(6):384-396.

58. Sipe JD, *et al.* (2012) Amyloid fibril protein nomenclature: 2012 recommendations from the Nomenclature Committee of the International Society of Amyloidosis. *Amyloid* 19(4):167-170.

59. Barnhart MM & Chapman MR (2006) Curli biogenesis and function. *Annu. Rev. Microbiol.* 60:131-147.

60. Fowler DM, *et al.* (2006) Functional amyloid formation within mammalian tissue. *PLoS Biol.* 4(1):e6.

61. Iconomidou VA, *et al.* (2006) Amyloid fibril formation propensity is inherent into the hexapeptide tandemly repeating sequence of the central domain of silkmoth chorion proteins of the A-family. *J. Struct. Biol.* 156(3):480-488.

62. Maji SK, *et al.* (2009) Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules. *Science* 325(5938):328-332.

63. Ando Y, *et al.* (2016) Effects of tafamidis treatment on transthyretin (TTR) stabilization, efficacy, and safety in Japanese patients with familial amyloid

polyneuropathy (TTR-FAP) with Val30Met and non-Val30Met: A phase III, open-label study. *J. Neurol. Sci.* 362:266-271.

64. Ellis JM & Fell MJ (2017) Current approaches to the treatment of Parkinson's Disease. *Bioorg. Med. Chem. Lett.* 27(18):4247-4255.

65. Lane CA, Hardy J, & Schott JM (2017) Alzheimer's disease. *Eur. J. Neurol.*

66. Jahn TR*, et al.* (2010) The common architecture of cross-beta amyloid. *J. Mol. Biol.* 395(4):717-727.

67. Eisenberg D & Jucker M (2012) The amyloid state of proteins in human diseases. *Cell* 148(6):1188-1203.

68. Kayed R & Lasagna-Reeves CA (2013) Molecular Mechanisms of Amyloid Oligomers Toxicity. *Journal of Alzheimers Disease* 33:S67-S78.

69. Rajasekhar K, Chakrabarti M, & Govindaraju T (2015) Function and toxicity of amyloid beta and recent therapeutic interventions targeting amyloid beta in Alzheimer's disease. *Chem. Commun. (Camb)* 51(70):13434-13450.

70. Milanesi L*, et al.* (2012) Direct three-dimensional visualization of membrane disruption by amyloid fibrils. *Proc Natl Acad Sci U S A* 109(50):20455-20460.

71. Sheynis T*, et al.* (2013) Aggregation modulators interfere with membrane interactions of beta2-microglobulin fibrils. *Biophys. J.* 105(3):745-755.

72. Cohen SI*, et al.* (2013) Proliferation of amyloid-beta42 aggregates occurs through a secondary nucleation mechanism. *Proc Natl Acad Sci U S A* 110(24):9758-9763.

73. Lee J, Culyba EK, Powers ET, & Kelly JW (2011) Amyloid-β forms fibrils by nucleated conformational conversion of oligomers. *Nat. Chem. Biol.* 7(9):602-609.

74. Bucciantini M*, et al.* (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416(6880):507-511.

75. Benilova I, Karran E, & De Strooper B (2012) The toxic Abeta oligomer and Alzheimer's disease: an emperor in need of clothes. *Nat. Neurosci.* 15(3):349-357.

76. Hayden EY & Teplow DB (2013) Amyloid beta-protein oligomers and Alzheimer's disease. *Alzheimers. Res. Ther.* 5(6):60.

77. Selkoe DJ & Hardy J (2016) The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* 8(6):595-608.

78. Haass C & Selkoe DJ (2007) Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nat. Rev. Mol. Cell Biol.* 8(2):101-112.

79. Kayed R*, et al.* (2003) Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* 300(5618):486-489.

80. Walsh DM*, et al.* (2002) Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation in vivo. *Nature* 416(6880):535-539.

81. Yang T, Li S, Xu H, Walsh DM, & Selkoe DJ (2017) Large Soluble Oligomers of Amyloid beta-Protein from Alzheimer Brain Are Far Less Neuroactive Than the Smaller Oligomers to Which They Dissociate. *J. Neurosci.* 37(1):152-163.

82. Prince M, Karagiannidou M, Guerchet M, Corras-Herrera A, & Knapp M (2016) World Alzheimer Report 2016: Improving healthcare for people living with dementia – coverage, quality and costs now and in the future. (World Health Organization, London).

83. Qiu C, Kivipelto M, & von Strauss E (2009) Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. *Dialogues Clin Neurosci* 11(2):111-128.

84. Trojanowski JQ & Lee VM (2000) "Fatal attractions" of proteins. A comprehensive hypothetical mechanism underlying Alzheimer's disease and other neurodegenerative disorders. *Ann. N. Y. Acad. Sci.* 924:62-67.

85. Braak H & Braak E (1991) [Morphological changes in the human cerebral cortex in dementia]. *J Hirnforsch* 32(3):277-282.

86. Masters CL*, et al.* (1985) Amyloid plaque core protein in Alzheimer disease and Down syndrome. *Proc Natl Acad Sci U S A* 82(12):4245-4249.

87. Gravina SA*, et al.* (1995) Amyloid beta protein (A beta) in Alzheimer's disease brain. Biochemical and immunocytochemical analysis with antibodies specific for forms ending at A beta 40 or A beta 42(43). *J. Biol. Chem.* 270(13):7013-7016.

88. Iwatsubo T*, et al.* (1994) Visualization of A beta 42(43) and A beta 40 in senile plaques with end-specific A beta monoclonals: evidence that an initially deposited species is A beta 42(43). *Neuron* 13(1):45-53.

89. Duff K*, et al.* (1996) Increased amyloid-beta42(43) in brains of mice expressing mutant presenilin 1. *Nature* 383(6602):710-713.

90. Kuperstein I*, et al.* (2010) Neurotoxicity of Alzheimer's disease Abeta peptides is induced by small changes in the Abeta42 to Abeta40 ratio. *EMBO J.* 29(19):3408-3420.

91. Danielsson J, Jarvet J, Damberg P, & Graslund A (2005) The Alzheimer beta-peptide shows temperature-dependent transitions between left-handed 3-helix, beta-strand and random coil secondary structures. *FEBS J.* 272(15):3938-3949.

92. Walti MA, Orts J, Vogeli B, Campioni S, & Riek R (2015) Solution NMR studies of recombinant Abeta(1-42): from the presence of a micellar entity to residual beta-sheet structure in the soluble species. *ChemBioChem* 16(4):659-669.

93. Yamaguchi T, Matsuzaki K, & Hoshino M (2011) Transient formation of intermediate conformational states of amyloid-beta peptide revealed by heteronuclear magnetic resonance spectroscopy. *FEBS Lett.* 585(7):1097-1102.

94. Roche J, Shen Y, Lee JH, Ying J, & Bax A (2016) Monomeric Abeta(1-40) and Abeta(1-42) Peptides in Solution Adopt Very Similar Ramachandran Map

Distributions That Closely Resemble Random Coil. *Biochemistry* 55(5):762-775.

95. Luhrs T, *et al.* (2005) 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *Proc Natl Acad Sci U S A* 102(48):17342-17347.

96. Petkova AT, Yau WM, & Tycko R (2006) Experimental constraints on quaternary structure in Alzheimer's beta-amyloid fibrils. *Biochemistry* 45(2):498-512.

97. Schmidt M, *et al.* (2015) Peptide dimer structure in an Abeta(1-42) fibril visualized with cryo-EM. *Proc Natl Acad Sci U S A* 112(38):11858-11863.

98. Xiao Y, *et al.* (2015) Abeta(1-42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* 22(6):499-505.

99. Kollmer M, *et al.* (2019) Cryo-EM structure and polymorphism of Abeta amyloid fibrils purified from Alzheimer's brain tissue. *Nat Commun* 10(1):4760.

100. Bertini I, Gonnelli L, Luchinat C, Mao J, & Nesi A (2011) A new structural model of Abeta40 fibrils. *J. Am. Chem. Soc.* 133(40):16013-16022.

101. Colvin MT, *et al.* (2016) Atomic Resolution Structure of Monomorphic Abeta42 Amyloid Fibrils. *J. Am. Chem. Soc.* 138(30):9663-9674.

102. Walti MA, *et al.* (2016) Atomic-resolution structure of a disease-relevant Abeta(1-42) amyloid fibril. *Proc Natl Acad Sci U S A* 113(34):E4976-4984.

103. Jarrett JT, Berger EP, & Lansbury PT, Jr. (1993) The C-terminus of the beta protein is critical in amyloidogenesis. *Ann. N. Y. Acad. Sci.* 695:144-148.

104. Meisl G, *et al.* (2014) Differences in nucleation behavior underlie the contrasting aggregation kinetics of the Abeta40 and Abeta42 peptides. *Proc Natl Acad Sci U S A* 111(26):9384-9389.

105. Cukalevski R, *et al.* (2015) The Abeta40 and Abeta42 peptides self-assemble into separate homomolecular fibrils in binary mixtures but cross-react during primary nucleation. *Chem Sci* 6(7):4215-4233.

106. Olzscha H, *et al.* (2011) Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell* 144(1):67-78.

107. Woerner AC, *et al.* (2016) Cytoplasmic protein aggregates interfere with nucleocytoplasmic transport of protein and RNA. *Science* 351(6269):173-176.

108. Dolfe L, Winblad B, Johansson J, & Presto J (2016) BRICHOS binds to a designed amyloid-forming beta-protein and reduces proteasomal inhibition and aggresome formation. *Biochem. J.* 473(2):167-178.

109. Vincenz-Donnelly L, *et al.* (2017) High capacity of the endoplasmic reticulum to prevent secretion and aggregation of amyloidogenic proteins. *EMBO J.*

110. Gosline JM, DeMont ME, & Denny MW (1986) The structure and properties of spider silk. *Endeavour* 10(1):37-43.

111. Swanson BO, Blackledge TA, Summers AP, & Hayashi CY (2006) Spider dragline silk: correlated and mosaic evolution in high-performance biological materials. *Evolution* 60(12):2539-2551.

112. Radtke C*, et al.* (2011) Spider silk constructs enhance axonal regeneration and remyelination in long nerve defects in sheep. *PLoS One* 6(2):e16990.

113. Allmeling C*, et al.* (2008) Spider silk fibres in artificial nerve constructs promote peripheral nerve regeneration. *Cell Prolif* 41(3):408-420.

114. Lewicka M, Hermanson O, & Rising AU (2012) Recombinant spider silk matrices for neural stem cell cultures. *Biomaterials* 33(31):7712-7717.

115. Wu S, Johansson J, Hovatta O, & Rising A (2016) Efficient passage of human pluripotent stem cells on spider silk matrices under xeno-free conditions. *Cell. Mol. Life Sci.* 73(7):1479-1488.

116. Rising A & Johansson J (2015) Toward spinning artificial spider silk. *Nat. Chem. Biol.* 11(5):309-315.

117. Rising A, Hjalm G, Engstrom W, & Johansson J (2006) N-terminal nonrepetitive domain common to dragline, flagelliform, and cylindriform spider silk proteins. *Biomacromolecules* 7(11):3120-3124.

118. Motriuk-Smith D, Smith A, Hayashi CY, & Lewis RV (2005) Analysis of the conserved N-terminal domains in major ampullate spider silk proteins. *Biomacromolecules* 6(6):3152-3159.

119. Kerkam K, Viney C, Kaplan D, & Lombardi S (1991) Liquid crystallinity of natural silk secretions. *Nature* 349(6310):596.

120. Knight DP & Vollrath F (2002) Spinning an elastic ribbon of spider silk. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357(1418):219-227.

121. Hijirida DH*, et al.* (1996) 13C NMR of Nephila clavipes major ampullate silk gland. *Biophys. J.* 71(6):3442-3447.

122. Chen X, Knight DP, & Vollrath F (2002) Rheological characterization of nephila spidroin solution. *Biomacromolecules* 3(4):644-648.

123. Jin HJ & Kaplan DL (2003) Mechanism of silk processing in insects and spiders. *Nature* 424(6952):1057-1061.

124. Lin Z, Huang W, Zhang J, Fan JS, & Yang D (2009) Solution structure of eggcase silk protein and its implications for silk fiber formation. *Proc Natl Acad Sci U S A* 106(22):8906-8911.

125. Parent LR*, et al.* (2018) Hierarchical spidroin micellar nanoparticles as the fundamental precursors of spider silks. *Proc Natl Acad Sci U S A* 115(45):11507-11512.

126. Vollrath F & Knight DP (2001) Liquid crystalline spinning of spider silk. *Nature* 410(6828):541-548.

127. Vollrath F & Knight DP (1999) Structure and function of the silk production pathway in the spider Nephila edulis. *Int. J. Biol. Macromol.* 24(2-3):243-249.

128. Andersson M*, et al.* (2014) Carbonic anhydrase generates CO2 and H+ that drive spider silk formation via opposite effects on the terminal domains. *PLoS Biol.* 12(8):e1001921.

129. Dicko C, Vollrath F, & Kenney JM (2004) Spider silk protein refolding is controlled by changing pH. *Biomacromolecules* 5(3):704-710.

130. Knight DP & Vollrath F (2001) Changes in element composition along the spinning duct in a Nephila spider. *Naturwissenschaften* 88(4):179-182.

131. Babb PL*, et al.* (2017) The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. *Nat. Genet.* 49(6):895-903.

132. Sanggaard KW*, et al.* (2014) Spider genomes provide insight into composition and evolution of venom and silk. *Nat Commun* 5:3765.

133. Hayashi CY, Shipley NH, & Lewis RV (1999) Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. *International Journal of Biological Macromolecules* 24(2):271-275.

134. Garb JE, Ayoub NA, & Hayashi CY (2010) Untangling spider silk evolution with spidroin terminal domains. *BMC Evol. Biol.* 10:243.

135. Challis RJ, Goodacre SL, & Hewitt GM (2006) Evolution of spider silks: conservation and diversification of the C-terminus. *Insect Mol. Biol.* 15(1):45-56.

136. Gao Z*, et al.* (2013) Structural characterization of minor ampullate spidroin domains and their distinct roles in fibroin solubility and fiber formation. *PLoS One* 8(2):e56142.

137. Hagn F*, et al.* (2010) A conserved spider silk domain acts as a molecular switch that controls fibre assembly. *Nature* 465(7295):239-242.

138. Wang S, Huang W, & Yang D (2014) Structure and function of C-terminal domain of aciniform spidroin. *Biomacromolecules* 15(2):468-477.

139. Sponner A*, et al.* (2005) The conserved C-termini contribute to the properties of spider silk fibroins. *Biochem. Biophys. Res. Commun.* 338(2):897-902.

140. Gauthier M, Leclerc J, Lefevre T, Gagne SM, & Auger M (2014) Effect of pH on the structure of the recombinant C-terminal domain of Nephila clavipes dragline silk protein. *Biomacromolecules* 15(12):4447-4454.

141. Askarieh G*, et al.* (2010) Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. *Nature* 465(7295):236-238.

142. Gaines WA, Sehorn MG, & Marcotte WR, Jr. (2010) Spidroin N-terminal domain promotes a pH-dependent association of silk proteins during self-assembly. *J. Biol. Chem.* 285(52):40745-40753.

143. Hagn F, Thamm C, Scheibel T, & Kessler H (2011) pH-dependent dimerization and salt-dependent stabilization of the N-terminal domain of spider dragline silk--implications for fiber formation. *Angew Chem Int Ed Engl* 50(1):310-313.

144. Landreh M*, et al.* (2010) A pH-dependent dimer lock in spider silk protein. *J. Mol. Biol.* 404(2):328-336.

145. Schwarze S, Zwettler FU, Johnson CM, & Neuweiler H (2013) The N-terminal domains of spider silk proteins assemble ultrafast and protected from charge screening. *Nat Commun* 4:2815.

146. Ries J, Schwarze S, Johnson CM, & Neuweiler H (2014) Microsecond folding and domain motions of a spider silk protein structural switch. *J. Am. Chem. Soc.* 136(49):17136-17144.

147. Heiby JC, Goretzki B, Johnson CM, Hellmich UA, & Neuweiler H (2019) Methionine in a protein hydrophobic core drives tight interactions required for assembly of spider silk. *Nat Commun* 10(1):4378.

148. Hedhammar M*, et al.* (2008) Structural properties of recombinant nonrepetitive and repetitive parts of major ampullate spidroin 1 from Euprosthenops australis: implications for fiber formation. *Biochemistry* 47(11):3407-3417.

149. Kaldmae M*, et al.* (2019) High intracellular stability of the spidroin N-terminal domain in spite of abundant amyloidogenic segments revealed by in-cell hydrogen/deuterium exchange mass spectrometry. *FEBS J.*

150. Wallace JA & Shen JK (2012) Unraveling A Trap-and-Trigger Mechanism in the pH-Sensitive Self-Assembly of Spider Silk Proteins. *J Phys Chem Lett* 3(5):658-662.

151. Jaudzems K*, et al.* (2012) pH-dependent dimerization of spider silk N-terminal domain requires relocation of a wedged tryptophan side chain. *J. Mol. Biol.* 422(4):477-487.

152. Atkison JH, Parnham S, Marcotte WR, Jr., & Olsen SK (2016) Crystal Structure of the Nephila clavipes Major Ampullate Spidroin 1A N-terminal Domain Reveals Plasticity at the Dimer Interface. *J. Biol. Chem.* 291(36):19006-19017.

153. Jiang W, Askarieh G, Shkumatov A, Hedhammar M, & Knight SD (2019) Structure of the N-terminal domain of Euprosthenops australis dragline silk suggests that conversion of spidroin dope to spider silk involves a conserved asymmetric dimer intermediate. *Acta Crystallogr D Struct Biol* 75(Pt 7):618-627.

154. Boehr DD, Nussinov R, & Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5(11):789-796.

155. Grunberg R, Leckner J, & Nilges M (2004) Complementarity of structure ensembles in protein-protein binding. *Structure* 12(12):2125-2136.

156. Wlodarski T & Zagrovic B (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proc Natl Acad Sci U S A* 106(46):19346-19351.

157. Sodoyer R (2004) Expression systems for the production of recombinant pharmaceuticals. *BioDrugs* 18(1):51-62.

158. Gellissen G*, et al.* (2005) New yeast expression platforms based on methylotrophic Hansenula polymorpha and Pichia pastoris and on dimorphic Arxula adeninivorans and Yarrowia lipolytica - a comparison. *FEMS Yeast Res.* 5(11):1079-1096.

159. Hamilton SR, *et al.* (2006) Humanization of yeast to produce complex terminally sialylated glycoproteins. *Science* 313(5792):1441-1443.

160. Korajczyk PJ, Zerbs S, & Collart FR (2001) Protein Production for Biotechnology. *eLS*:1-9.

161. Demain AL & Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. *Biotechnol. Adv.* 27(3):297-306.

162. Murthy TV, *et al.* (2004) Bacterial cell-free system for high-throughput protein expression and a comparative analysis of Escherichia coli cell-free and whole cell expression systems. *Protein Expr Purif* 36(2):217-225.

163. Sezonov G, Joseleau-Petit D, & D'Ari R (2007) Escherichia coli physiology in Luria-Bertani broth. *J. Bacteriol.* 189(23):8746-8749.

164. Olins PO & Lee SC (1993) Recent advances in heterologous gene expression in Escherichia coli. *Curr. Opin. Biotechnol.* 4(5):520-525.

165. Chambers SP, Austen DA, Fulghum JR, & Kim WM (2004) High-throughput screening for soluble recombinant expressed kinases in Escherichia coli and insect cells. *Protein Expr. Purif.* 36(1):40-47.

166. Marblestone JG, *et al.* (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci.* 15(1):182-189.

167. Carrio M, Gonzalez-Montalban N, Vera A, Villaverde A, & Ventura S (2005) Amyloid-like properties of bacterial inclusion bodies. *J. Mol. Biol.* 347(5):1025-1037.

168. Wang L (2009) Towards revealing the structure of bacterial inclusion bodies. *Prion* 3(3):139-145.

169. Wang L, Maji SK, Sawaya MR, Eisenberg D, & Riek R (2008) Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol.* 6(8):e195.

170. Mitraki A & King J (1989) Protein folding intermediates and inclusion body formation. *Nat. Biotechnol.* 7(7):690-697.

171. Garcia-Fruitos E, *et al.* (2012) Bacterial inclusion bodies: making gold from waste. *Trends Biotechnol.* 30(2):65-70.

172. Kane JF & Hartley DL (1988) Formation of recombinant protein inclusion bodies in Escherichia coli. *Trends Biotechnol.* 6(5):95-101.

173. Marston FA (1986) The purification of eukaryotic polypeptides synthesized in Escherichia coli. *Biochem. J.* 240(1):1-12.

174. Valax P & Georgiou G (1993) Molecular characterization of beta-lactamase inclusion bodies produced in Escherichia coli. 1. Composition. *Biotechnol. Prog.* 9(5):539-547.

175. Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, & Hard T (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein Sci.* 11(2):313-321.

176. Kataeva I, *et al.* (2005) Improving solubility of Shewanella oneidensis MR-1 and Clostridium thermocellum JW-20 proteins expressed into Escherichia coli. *J Proteome Res* 4(6):1942-1951.

177. Turner P, Holst O, & Karlsson EN (2005) Optimized expression of soluble cyclomaltodextrinase of thermophilic origin in Escherichia coli by using a soluble fusion-tag and by tuning of inducer concentration. *Protein Expr. Purif.* 39(1):54-60.

178. Baneyx F & Palumbo JL (2003) Improving heterologous protein folding via molecular chaperone and foldase co-expression. *Methods Mol. Biol*. 205:171-197.

179. Kameoka D, Masuzaki E, Ueda T, & Imoto T (2007) Effect of buffer species on the unfolding and the aggregation of humanized IgG. *J Biochem* 142(3):383-391.

180. Katayama DS, *et al.* (2006) Effect of buffer species on the thermally induced aggregation of interferon-tau. *J. Pharm. Sci.* 95(6):1212-1226.

181. Hauptmann A, *et al.* (2018) Impact of Buffer, Protein Concentration and Sucrose Addition on the Aggregation and Particle Formation during Freezing and Thawing. *Pharm Res* 35(5):101.

182. Meliga SC, Farrugia W, Ramsland PA, & Falconer RJ (2013) Cold-induced precipitation of a monoclonal IgM: a negative activation enthalpy reaction. *J Phys Chem B* 117(2):490-494.

183. Castellanos MM, McAuley A, & Curtis JE (2017) Investigating Structure and Dynamics of Proteins in Amorphous Phases Using Neutron Scattering. *Comput Struct Biotechnol J* 15:117-130.

184. Pikal-Cleland KA, Cleland JL, Anchordoquy TJ, & Carpenter JF (2002) Effect of glycine on pH changes and protein stability during freeze-thawing in phosphate buffer systems. *J. Pharm. Sci.* 91(9):1969-1979.

185. Cicerone MT, Pikal MJ, & Qian KK (2015) Stabilization of proteins in solid form. *Adv Drug Deliv Rev* 93:14-24.

186. Roughton BC, Iyer LK, Bertelsen E, Topp EM, & Camarda KV (2013) Protein aggregation and lyophilization: Protein structural descriptors as predictors of aggregation propensity. *Comput. Chem. Eng.* 58(2013):369-377.

187. LaVallie ER & McCoy JM (1995) Gene fusion expression systems in Escherichia coli. *Curr. Opin. Biotechnol.* 6(5):501-506.

188. Waugh DS (2005) Making the most of affinity tags. *Trends Biotechnol.* 23(6):316-320.

189. Esposito D & Chatterjee DK (2006) Enhancement of soluble protein expression through the use of fusion tags. *Curr. Opin. Biotechnol.* 17(4):353-358.

190. Kimple ME, Brill AL, & Pasker RL (2013) Overview of affinity tags for protein purification. *Curr Protoc Protein Sci* 73:9 9 1-9 9 23.

191. Ha KS, Toulokhonov I, Vassylyev DG, & Landick R (2010) The NusA N-terminal domain is necessary and sufficient for enhancement of transcriptional

pausing via interaction with the RNA exit channel of RNA polymerase. *J. Mol. Biol.* 401(5):708-725.

192. Davis GD, Elisee C, Newham DM, & Harrison RG (1999) New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol. Bioeng.* 65(4):382-388.

193. Lavallie ER*, et al.* (1993) A Thioredoxin Gene Fusion Expression System That Circumvents Inclusion Body Formation in the Escherichia-Coli Cytoplasm. *Bio-Technology* 11(2):187-193.

194. Kapust RB & Waugh DS (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8(8):1668-1674.

195. Sachdev D & Chirgwin JM (1999) Properties of soluble fusions between mammalian aspartic proteinases and bacterial maltose-binding protein. *J. Protein Chem.* 18(1):127-136.

196. Raran-Kurussi S & Waugh DS (2012) The ability to enhance the solubility of its fusion partners is an intrinsic property of maltose-binding protein but their folding is either spontaneous or chaperone-mediated. *PLoS One* 7(11):e49589.

197. di Guana C, Lib P, Riggsa PD, & Inouyeb H (1988) Vectors that facilitate the expression and purification of foreign peptides in Escherichia coli by fusion to maltose-binding protein. *Gene* 67(1):21-30.

198. Smith DB & Johnson KS (1988) Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* 67(1):31-40.

199. Butt TR, Edavettal SC, Hall JP, & Mattern MR (2005) SUMO fusion technology for difficult-to-express proteins. *Protein Expression Purif.* 43(1):1-9.

200. Cheng Y & Patel DJ (2004) An efficient system for small protein expression and refolding. *Biochem. Biophys. Res. Commun.* 317(2):401-405.

201. Malakhov MP*, et al.* (2004) SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J Struct Funct Genomics* 5(1-2):75-86.

202. Structural Genomics C*, et al.* (2008) Protein production and purification. *Nat. Methods* 5(2):135-146.

203. Janson J-C (2012) *Protein purification: principles, high resolution methods, and applications* (John Wiley & Sons).

204. Nilapwar SM, Nardelli M, Westerhoff HV, & Verma M (2011) Absorption spectroscopy. *Methods Enzymol.*, (Elsevier), Vol 500, pp 59-75.

205. Lakowicz JR (2013) *Principles of fluorescence spectroscopy* (Springer Science & Business Media).

206. Fasman GD (2013) *Circular dichroism and the conformational analysis of biomolecules* (Springer Science & Business Media).

207. Cavanagh J, Fairbrother WJ, Palmer III AG, & Skelton NJ (1995) *Protein NMR spectroscopy: principles and practice* (Elsevier).

208. Hoffman ED & Stroobant V (2007) Mass spectrometry: principles and applications. *West Sussex: John Wiley & Sons, Bruxellas, Bélgica* 1(2):85.

209. Masson GR, *et al.* (2019) Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* 16(7):595.

210. Gade Malmos K, *et al.* (2017) ThT 101: a primer on the use of thioflavin T to investigate amyloid formation. *Amyloid* 24(1):1-16.

211. Foelix R (2011) *Biology of spiders* (OUP USA).

212. Otikovs M, *et al.* (2015) Diversified Structural Basis of a Conserved Molecular Mechanism for pH-Dependent Dimerization in Spider Silk N-Terminal Domains. *ChemBioChem* 16(12):1720-1724.

213. Bauer J, *et al.* (2016) Acidic Residues Control the Dimerization of the N-terminal Domain of Black Widow Spiders' Major Ampullate Spidroin 1. *Sci Rep* 6:34442.

214. Reid KB (1998) Interactions of surfactant protein D with pathogens, allergens and phagocytes. *Biochim. Biophys. Acta* 1408(2-3):290-295.

215. Madan T, *et al.* (2001) Surfactant proteins A and D protect mice against pulmonary hypersensitivity induced by Aspergillus fumigatus antigens and allergens. *J. Clin. Invest.* 107(4):467-475.

216. Almlen A, *et al.* (2008) Surfactant proteins B and C are both necessary for alveolar stability at end expiration in premature rabbits with respiratory distress syndrome. *J Appl Physiol (1985)* 104(4):1101-1108.

217. Almlen A, *et al.* (2010) Synthetic surfactant based on analogues of SP-B and SP-C is superior to single-peptide surfactants in ventilated premature rabbits. *Neonatology* 98(1):91-99.

218. Curstedt T, Jornvall H, Robertson B, Bergman T, & Berggren P (1987) Two hydrophobic low-molecular-mass protein fractions of pulmonary surfactant. Characterization and biophysical activity. *Eur. J. Biochem.* 168(2):255-262.

219. Whitsett JA & Weaver TE (2002) Hydrophobic surfactant proteins in lung function and disease. *N Engl J Med* 347(26):2141-2148.

220. Kallberg Y, Gustafsson M, Persson B, Thyberg J, & Johansson J (2001) Prediction of amyloid fibril-forming proteins. *J. Biol. Chem.* 276(16):12945-12950.

221. Nilsson G, *et al.* (1998) Synthetic peptide-containing surfactants--evaluation of transmembrane versus amphipathic helices and surfactant protein C poly-valyl to poly-leucyl substitution. *Eur. J. Biochem.* 255(1):116-124.

222. Sato A & Ikegami M (2012) SP-B and SP-C containing new synthetic surfactant for treatment of extremely immature lamb lung. *PLoS One* 7(7):e39392.

223. Seehase M, *et al.* (2012) New surfactant with SP-B and C analogs gives survival benefit after inactivation in preterm lambs. *PLoS One* 7(10):e47631.

224. Sweet DG, *et al.* (2017) A first-in-human clinical study of a new SP-B and SP-C enriched synthetic surfactant (CHF5633) in preterm babies with respiratory distress syndrome. *Arch Dis Child Fetal Neonatal Ed*.

225. Cochrane CG & Revak SD (1991) Pulmonary surfactant protein B (SP-B): structure-function relationships. *Science* 254(5031):566-568.

226. Revak SD, *et al.* (1996) Efficacy of synthetic peptide-containing surfactant in the treatment of respiratory distress syndrome in preterm infant rhesus monkeys. *Pediatr Res* 39(4 Pt 1):715-724.

227. Moya FR, *et al.* (2005) A multicenter, randomized, masked, comparison trial of lucinactant, colfosceril palmitate, and beractant for the prevention of respiratory distress syndrome among very preterm infants. *Pediatrics* 115(4):1018-1029.

228. Sinha SK, *et al.* (2005) A multicenter, randomized, controlled trial of lucinactant versus poractant alfa among very premature infants at high risk for respiratory distress syndrome. *Pediatrics* 115(4):1030-1038.

229. Mack L, Brill B, Delis N, & Groner B (2014) Endotoxin depletion of recombinant protein preparations through their preferential binding to histidine tags. *Anal. Biochem.* 466:83-88.

230. Basabe-Burgos O, *et al.* (2019) Natural Derived Surfactant Preparation As a Carrier of Polymyxin E for Treatment of Pseudomonas aeruginosa Pneumonia in a Near-Term Rabbit Model. *J Aerosol Med Pulm Drug Deliv* 32(2):110-118.

231. Willander H, *et al.* (2012) High-resolution structure of a BRICHOS domain and its implications for anti-amyloid chaperone activity on lung surfactant protein C. *Proc Natl Acad Sci U S A* 109(7):2325-2329.

232. Chen G, *et al.* (2017) Bri2 BRICHOS client specificity and chaperone activity are governed by assembly state. *Nat Commun* 8(1):2081.

233. Cohen SI, *et al.* (2015) A molecular chaperone breaks the catalytic cycle that generates toxic Abeta oligomers. *Nat. Struct. Mol. Biol.* 22(3):207-213.

234. Willander H, *et al.* (2012) BRICHOS domains efficiently delay fibrillation of amyloid beta-peptide. *J. Biol. Chem.* 287(37):31608-31617.

235. Hermansson E, *et al.* (2014) The chaperone domain BRICHOS prevents CNS toxicity of amyloid-beta peptide in Drosophila melanogaster. *Dis Model Mech* 7(6):659-665.

236. Poska H, *et al.* (2016) Dementia-related Bri2 BRICHOS is a versatile molecular chaperone that efficiently inhibits Abeta42 toxicity in Drosophila. *Biochem. J.* 473(20):3683-3704.

237. Klement K, *et al.* (2007) Effect of different salt ions on the propensity of aggregation and on the structure of Alzheimer's abeta(1-40) amyloid fibrils. *J. Mol. Biol.* 373(5):1321-1333.

238. Diaz-Espinoza R, Nova E, & Monasterio O (2017) Overcoming electrostatic repulsions during amyloid assembly: Effect of pH and interaction with divalent metals using model peptides. *Arch Biochem Biophys* 621:46-53.

239. Lowe R, Pountney DL, Jensen PH, Gai WP, & Voelcker NH (2004) Calcium(II) selectively induces alpha-synuclein annular oligomers via interaction with the C-terminal domain. *Protein Sci.* 13(12):3245-3252.

240. Abelein A, Jarvet J, Barth A, Graslund A, & Danielsson J (2016) Ionic Strength Modulation of the Free Energy Landscape of Abeta40 Peptide Fibril Formation. *J. Am. Chem. Soc.* 138(21):6893-6902.

241. Finder VH, Vodopivec I, Nitsch RM, & Glockshuber R (2010) The recombinant amyloid-beta peptide Abeta1-42 aggregates faster and is more neurotoxic than synthetic Abeta1-42. *J. Mol. Biol.* 396(1):9-18.

242. Garai K, Crick SL, Mustafi SM, & Frieden C (2009) Expression and purification of amyloid-beta peptides from Escherichia coli. *Protein Expr Purif* 66(1):107-112.

243. Long F, Cho W, & Ishii Y (2011) Expression and purification of 15N- and 13C-isotope labeled 40-residue human Alzheimer's beta-amyloid peptide for NMR-based structural analysis. *Protein Expr Purif* 79(1):16-24.

244. Lindberg DJ & Esbjorner EK (2016) Detection of amyloid-beta fibrils using the DNA-intercalating dye YOYO-1: Binding mode and fibril formation kinetics. *Biochem. Biophys. Res. Commun.* 469(2):313-318.

245. Lindberg DJ, Wesen E, Bjorkeroth J, Rocha S, & Esbjorner EK (2017) Lipid membranes catalyse the fibril formation of the amyloid-beta (1-42) peptide through lipid-fibril interactions that reinforce secondary pathways. *Biochim Biophys Acta Biomembr* 1859(10):1921-1929.

246. Lawrence MS, Phillips KJ, & Liu DR (2007) Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* 129(33):10110-10112.

247. Zhang Y*, et al.* (1998) Expression of eukaryotic proteins in soluble form in Escherichia coli. *Protein Expr Purif* 12(2):159-165.

248. Chen G*, et al.* (2012) Full-length minor ampullate spidroin gene sequence. *PLoS One* 7(12):e52293.